# UCONGA: Universal Conformer Generation and Analysis

*A thesis submitted in partial fulfillment of the requirements for the degree of*

*Doctor of Philosophy in Chemistry*

*at the*

*University of Canterbury, Christchurch, New Zealand*

**Nathaniel Gunby**

**2016**

## Table of Contents

## Acknowledgements

First and foremost, I would like to thank my supervisory team for the last three and a quarter years, Drs. Sarah Masters and Deborah Crittenden, for their useful advice and many hours proofreading this thesis. To Sarah, thanks in particular for encouraging me to extend myself – without being advised 'If you don't ask, you don't get' I would not have travelled to conferences on three continents over the course of my PhD. To Deborah, thanks in particular for knowing when side projects were productive, when they should have side-projects of their own explored, and when they should be abandoned.

I would also like to thank Susan Krumdieck and her research group in the Department of Mechanical Engineering, for getting me interested in aluminum isopropoxide and conformer generation in the first place. While this research wasn't the intended outcome of our collaboration, I think we did both get something out of it.

The physical chemistry research group while I've been here – Alex, Andrew, Andrew, Chris, Chris, Dinga, Heather, Marat and Sandra – have played a big role in getting me through this PhD journey. Thanks for the cake and discussions both enjoyable and useful. Thanks especially to Sandra for staying calm and positive while we were stranded overnight in San Francisco airport by bad weather.

## Abstract

This work presents a solution to the problem of adequately sampling large conformational spaces computationally using a deterministic method that is guaranteed to identify all of the sterically-allowed molecular conformations to within a given bond rotation resolution. This involved development of new conformer generation software called UCONGA and its use, along with other known computational methods, to study various systems with interesting conformational properties.

This thesis presents details of the new conformer ensemble generation method UCONGA, and the tools developed simultaneously alongside UCONGA to analyze the generated conformer ensembles. These analysis tools aim to find clusters of similar conformers so that representative or unusual ones can be selected for further study, and can be used on conformer ensembles generated through other methods as well as those generated by UCONGA. It then discusses the extension of UCONGA with a divide-and-conquer algorithm to improve its performance with increasing molecular size. Initially, the suitability of various computational chemistry methods for studying conformer ensembles is tested. A series of case studies are then presented sampling different challenges for conformer generation methods. This first case study examines extremely sterically crowded molecules with few stable conformers, which often adopt unusual dihedral angles to avoid steric clashing. The second case study is on molecules bound to surfaces, which require different metrics for the conformational properties, and are in a different chemical environment than the other case studies. The final case study is on highly flexible molecules with large conformer ensembles.

## Abbreviations

| | |
|---|---|
| B3LYP | The combination of Becke's 3-parameter electron-exchange functional with the Lee-Yang-Parr electron correlation functional |
| DFT | Density Functional Theory |
| GED | Gas-phase Electron Diffraction |
| HF | Hartree-Fock theory |
| M06 | The 2006 Minnesota University density functional of Truhlar *et al.* |
| MP2 | Moller-Plesset perturbation theory truncated at the second order |
| RMSD | Root-mean-squared deviation in atomic positions |
| UCONGA | Universal Conformer GENeration and Analysis |

# Chapter 1

## General overview

## 1.1 The conformational sampling problem

Molecular structure – that is, the precise three-dimensional arrangement of atoms in a molecule – is important because it affects molecular activity, such as the ability to act as a catalyst or drug. Hundreds of thousands of papers on structure-activity relationships in the scientific literature [1] attest to the importance of determining molecular structure *en route* to predicting molecular properties. For example, molecular conformation affects the strength of intermolecular interactions, so the first step in any computational drug design and docking study involves generating an ensemble of possible conformers for the ligand before assessing the ability of that ligand to bind to a target protein [2]. Molecular conformation also affects electronic properties, as it affects the degree of orbital overlap. For instance, the fluorescence wavelength of para-*N*,*N*-dimethylamino-benzonitrile is conformer-dependent [3-4].

The structure of a molecule can be written in terms of Cartesian coordinates, describing the positions of the atoms, or in internal coordinates, relating the positions of the atoms to each other through bond lengths, bond angles and torsion angles. Each unique set of Cartesian or internal coordinates is referred to as a molecular conformation and stable conformers are identified when any change in coordinate values results in an increase in energy. For a given molecule, each bond length and bond angle typically has only one stable value. However, bond rotation can be more facile, with multiple possible rotamers accessible at room temperature. Therefore, more than one conformer is generally involved in any chemical process, so the entire ensemble, not just the lowest-energy conformer, must be found. Finding the conformers of a molecule becomes increasingly difficult as its size, specifically the number of rotatable bonds, increases, since the number of possible conformers grows exponentially with the number of rotatable bonds to be interrogated.

Due to the importance of the problem, various computational techniques to generate conformer ensembles have already been developed. These fall into four classes: deterministic, reduced-dimensional, knowledge-based and stochastic. The strengths and weaknesses of each of these will now be discussed in more detail.

The conceptually simplest conformer ensemble generation technique is the *ab initio* potential energy surface search, a deterministic (for small molecules) or reduced-

dimensional (for larger molecules) process where one or more bonds are systematically rotated around in specified increments and the energy associated with the remainder of the system minimized using a constrained geometry optimization process. However, due to the computational resources required for these calculations, no more than two bonds at a time can be rotated, making it reduced-dimensional instead of deterministic for molecules with more than two bonds that can rotate. This makes it incapable of accurately generating conformers for systems where three or more bonds can meaningfully rotate simultaneously, such as large branched systems. Additionally, the results of the optimization over the remaining coordinates are dependent on the starting conformation, so there is no guarantee that all unique stable conformers will be found. Therefore, these methods are the best option for molecules with fewer rotatable bonds but unusual bonding patterns and heavier atoms. Compounds containing silicon, phosphorus, or sulfur are often best described using these methods

There are a number of knowledge-based methods such as ALFA [5] and CONFAB [6] designed to generate ligand conformer ensembles for protein-ligand docking. These methods use forcefields that are only parameterized for a common subset of the chemical elements, typically C, H, N, O, P, S and the halogens. This parameterization limits the range of these methods. Some protein-bound ligand structures containing elements they are not parameterized for have had to be excluded from their testing [7-8] and they are inapplicable to inorganic molecules and complexes. These methods are best suited to moderately-sized organic molecules of no more than ten rotatable bonds.

Stochastic methods, such as genetic algorithms [7] and Monte Carlo methods [9], can be set to only generate a certain number of conformers. This allows their runtime to be roughly independent of molecular size, but they are not deterministic and may therefore fail to generate the lowest-energy conformer and may also fail to generate a representative conformer ensemble. These methods are the best option for molecules with 10-20 rotatable bonds.

Between the four classes of method, organic molecules of up to 10 rotatable bonds can have conformer ensembles generated in a deterministic fashion, but there is no such method for inorganic molecules with more than 2 rotatable bonds. This thesis describes the creation of a universal deterministic conformer generation method, called UCONGA, which can generate deterministic conformer ensembles for inorganic

molecules in this size range. The name UCONGA stands for Universal CONformer Generation and Analysis. This method development has been coupled with the development of tools for analyzing conformer ensembles. These tools can help extract chemical meaning from conformer ensembles generated by UCONGA or other methods for communication with experimental colleagues or planning further, more focused, computational studies. This method development is described in more detail in Chapter 2.

## 1.2 Case Studies

For the UCONGA method to live up to the universal title, it must be tested against a wide range of molecules. Different types of system present different problems for conformer identification. Extremely sterically crowded molecules have few conformers even with many rotatable bonds, but the conformers often adopt unusual torsion angles. For example, the lowest-energy conformers of 1,1,2,2-tetra-*tert*-butyl disilane [10] and 1,1,2-tri-*tert*-butyl disilane [11] both feature *tert*-butyl groups eclipsed with hydrogen atoms when viewed down the central Si-Si bond to avoid gauche interactions between the *tert*-butyl groups. Less crowded molecules with many rotatable bonds present a different challenge, as they can have very large conformer ensembles. Much of the challenge in creating a universal conformer ensemble generation method comes from the need to build in enough chemical knowledge to avoid taking too much time and generating overly large ensembles full of high-energy conformers for molecules with many rotatable bonds, while not making the method overly specialized and incapable of finding conformers for the more unusual cases.

The environment of the molecules presents a further challenge. Chemical processes usually occur in the presence of other molecules, such as in solution, at a surface, or in the binding pocket of a protein. These affect the adopted conformers as well. A universal conformer generation method must be able to generate conformers that can exist in all of these environments.

While UCONGA aims to be a universal conformer generation method, it cannot provide energetic information about the conformers it generates. Universal conformer energies can only come from *ab initio* methods. These methods lie along a spectrum from fast and approximate to slow and accurate and it is important to make the best possible trade-off

between accuracy and computational cost, particularly when generating optimized geometries and energies across large conformer ensembles. Therefore, before in-depth case studies are performed, the ability of a collection of these computational methods to correctly reproduce and energy-rank the conformers of a subset of the molecules to be studied will be tested.

Three sets of case studies have been chosen to investigate and demonstrate the universality of UCONGA. The first set of case studies, presented in Chapter 5, will focus on the challenges associated with generating conformers of highly crowded molecules, focusing particularly on inorganic gas-phase systems for which experimental reference data are available. The second, presented in Chapter 6, will focus on studying moderately-sized molecules (3-5 rotatable bonds) attached to a surface. These are interesting due to the uniqueness of the environment and its effect on the analysis methods of UCONGA. A molecule bound to a surface is not free to rotate in all dimensions, but can only rotate around the axis perpendicular to the surface. The final set of case studies, presented in Chapter 7, will focus on the challenges posed by highly flexible molecules. This will test the ability of UCONGA to generate diverse conformer ensembles containing experimentally relevant conformers across different chemical environments and a range of molecular structures. This will involve both protein-bound ligands, which are organic and relatively linear, and gaseous dimeric aluminum isopropoxide, which is inorganic and highly branched. The combination of these test cases – sterically hindered gas-phase inorganic molecules, moderately flexible molecules at a surface, flexible organic molecules bound within a crystal and flexible gas-phase inorganic molecules – provide a diverse set of molecules, varying from highly crowded to highly flexible, with which to test the universality of UCONGA.

## 1.3 References


1) 175,832 search results on www.scopus.com for 'structure-activity relationship' on 22/02/2016, limited to biochemistry, pharmacology, chemistry, medicine, immunology, chemical engineering and neurology.
2) Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discov. Today Technol.* **2010**, *7* (4), e245–e253.
3) Rotkiewicz, K.; Grellmann, K. H.; Grabowski, Z. R. Reinterpretation of the Anomalous Fluorescense of *p*-*N*,*N*-Dimethylamino-Benzonitrile. *Chem. Phys. Lett.* **1973**, *19* (3), 315–318.
4) Rettig, W. Charge Separation in Excited States of Decoupled Systems – TICT Compounds and Implications Regarding the Development of New Laser Dyes and the Primary Processes of Vision and Photosynthesis. *Angew. Chemie-International Ed. English* **1986**, *25* (11), 971–988.
5) Klett, J.; Cortés-Cabrera, Á.; Gil-Redondo, R.; Gago, F.; Morreale, A. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.* **2014**, *54*, 314–323.
6) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab – Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.
7) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474.
8) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good are They? *J. Chem. Inf. Model.* **2012**, *52* (5), 1146–1158.
9) Miteva, M. A; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W622–W627.
10) Hinchley, S. L.; Robertson, H. E.; Parkin, A.; Rankin, D. W. H.; Tekautz, G.; Hassler, K. Molecular Structure of 1,1,2,2-Tetra-*tert*-butyldisilane: Unusual Structural Motifs in Sterically Crowded Disilanes. *Dalton Trans.* **2004**, (5), 759–766.
11) Hinchley, S. L.; Smart, B. A.; Morrison, C. A.; Robertson, H. E.; Rankin, D. W. H.; Zink, R.; Hassler, K. 1, 1, 2-Tri-*tert*-butyldisilane, $Bu^t_2HSiSiH_2Bu^t$: Vibrational Spectra and Molecular Structure in the Gas Phase by Electron Diffraction and *Ab Initio* Calculations. *J. Chem. Soc., Dalton Trans.* **1999**, 2303–2310.

# Chapter 2

## Development and implementation of a method for Universal CONformer Generation and Analysis

## 2.1 Introduction

Generating an ensemble of low-energy conformers is an important first step in many computational chemistry studies. For example, all protein-ligand docking methods require an ensemble of conformers for the ligand [1]. As such, there are many programs available to create such an ensemble, including ALFA [2], Balloon [3], Confab [4], ConfGen [5] and OMEGA [6]. Their specialized design lacks versatility, which can be problematic for their intended use in protein-ligand docking and prevents their adoption in other fields where computational conformer location is important, including gas-phase structural chemistry [7].

These methods are overspecialized because they all use one of the many forcefields designed for organic and biological chemistry. However, these forcefields are only parameterized for a common subset of the chemical elements. In testing these methods, compounds had to be excluded from the test set simply because they contained elements that the forcefield was not parameterized for [3, 8]. This is a problem for computational drug design, as elements that many forcefields are poorly parameterized for, including silicon and arsenic, have been used in active drugs [9-13]. It is a more significant hindrance for structural chemistry, where 'uncommon' elements are, in fact, common.

The Universal Conformer Generation and Analysis (UCONGA) method has been designed to meet this need for a more general conformer generation technique. UCONGA is a conformer ensemble generation method that does not rely on forcefields. It is nearly parameter-free, using only van der Waals radii. Therefore conformer ensembles can be generated for any molecule, including molecules with non-organic elements and unusual structural features that are not well-described by common forcefields. This includes ring conformer generation without the use of a library of known ring conformers. The UCONGA method is also capable of analyzing conformer ensembles to find a representative set of conformers for further study if the generated conformer ensemble is too large.

This chapter provides a discussion of how the UCONGA method works and how it was implemented, and provides some benchmarking data showing its abilities and limitations.

## 2.2 The UCONGA method

### 2.2.1: Overview

There are three main steps to the UCONGA algorithm: pre-analysis of the molecular structure, generation of trial conformers and screening of trial conformers. In the pre-analysis step, information is derived about the symmetry and connectivity of the molecule that helps make conformer generation more efficient. In the trial conformation generation step, unique conformations are generated without any regard for their stability. Finally, the screening step removes unstable trial conformers from the generated ensemble. This is summarized in Figure 2.1. These will now be discussed in greater detail.

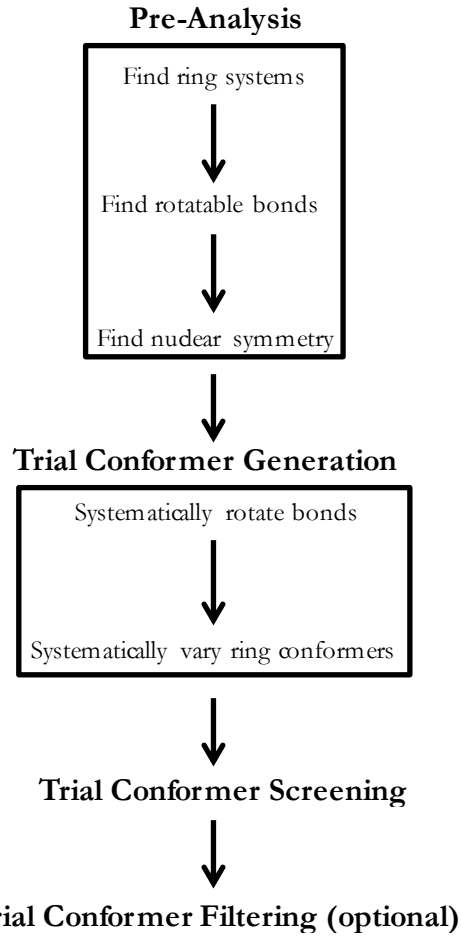



**Figure 2.1** A flowchart showing the main steps of conformer generation with UCONGA as well as the substeps of the more complicated steps.

### 2.2.2: Pre-analysis

Pre-analysis of the molecule has three aims: finding the ring systems, identifying the rotatable bonds and determining nuclear permutational symmetry i.e. the ways in which atoms can have their labels interchanged without breaking connectivity. First, rings are

identified by analyzing atomic connectivities. During subsequent bond rotations, bridged and fused rings are treated as a single unit because their conformational changes are coupled. Spiro rings, however, are treated independently as their conformational changes are not necessarily coupled (Figure 2.2).

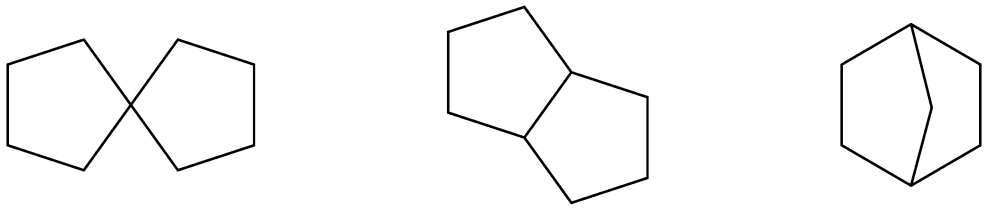


**Figure 2.2** Spiro rings (left) can undergo independent conformational change, while fused (middle) and bridged (right) rings cannot.

Next, rotatable bonds are defined as single bonds that are not part of a ring system and have at least one non-hydrogen substituent attached to each end. Finally, nuclear permutational symmetry is identified, using the modified Morgan algorithm[14] to find atoms in identical chemical environments. The basic Morgan algorithm finds the chemical environments of atoms by assigning heavy valence identifiers to each atom where the heavy valence is the number of bonded non-hydrogen atoms. These identifiers are iteratively updated by adding all neighboring values at each step until the number of different values does not change. Each atom with a different identifier at the end of this process is in a different chemical environment. The modified Morgan algorithm also performs checks that R/S or E/Z stereocentres are not being treated identically and in this work it has been further modified to account for parastereocenters [15].

### 2.2.3: Trial conformer generation

Maintaining universality while efficiently generating trial conformers is a challenge. While many other conformer ensemble generation methods rely on a rules-driven approach using a list of preferred torsion angles, UCONGA cannot use this approach, as it will not find conformers for molecules containing high levels of steric crowding or unusual bonding patterns. UCONGA instead systematically rotates all rotatable bonds, as identified during pre-analysis, in a stepwise fashion.

To avoid locating multiple conformations in a given basin on the potential energy surface, a multi-step process is used. A first scan is performed with a relatively large step size and then finer-grained searches are performed in areas of the potential energy surface where no conformer has yet been found. The number of trial conformers is

further reduced by decreasing the maximum torsion angle for rotation about symmetrical bonds, which include symmetric rotors and bonds that are equivalent under nuclear permutational symmetry, as illustrated in Figure 2.3 and discussed below.

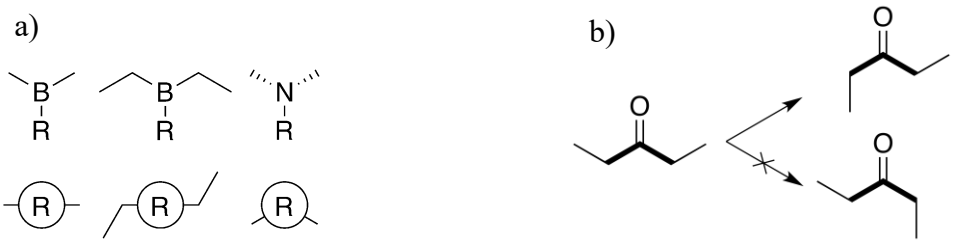



**Figure 2.3a)** Left: a dimethylboryl group is a symmetric rotor of order 2. Middle: a diethylboryl group is not a symmetric rotor because the ethyl groups connected to the boron atom contain rotatable bonds. Right: a dimethylamino group is not a symmetric rotor because the methyl groups are not equidistant; i.e, they do not evenly divide the circle in a Newman projection.
**Figure 2.3b)** Left: An allowable conformer for pentan-3-one with torsion angles around symmetry-equivalent rotatable bonds (indicated in bold) identical. Right top: another allowable conformer for pentan-3-one, with the torsion around the second, right-hand equivalent bond less than that of the first. Right bottom: a forbidden conformer identical to the right top conformer but with the torsion angle around the second equivalent bond greater than that around the first.

A symmetric rotor is formed by an atom attached to terminal non-rotatable groups that are equidistant from each other and in the same symmetry class, as illustrated in Figure 2.3 a. In these cases, the maximum torsion angle around the bond to the symmetric rotor is divided by the order of symmetry of the rotor. In this context, the order of symmetry of the rotor is equivalent to the number of attached terminal groups.

Rotations around equivalent bonds lead to automorphically equivalent conformers, as shown on the right hand side of Figure 2.3 b. A bond is defined as equivalent to another bond if the symmetry classes of the atoms involved are equivalent and the two bonds share a common atom. At no stage in the conformer generation process may the torsion angle of the second equivalent bond exceed that of the first. This ensures that all conformers generated are distinct.

Because the rotations of bonds that are in rings are concerted, rotating them in this straightforward fashion is impractical. Instead, UCONGA generates ring conformers using the flip-of-fragments method [16]. This is a two-step process. In the first step, a two-atom section of the ring with all substituents is reflected through the plane defined

by its junction with the rest of the ring. In the second step, each atom in the ring that moved in the first step and all its substituents are reflected in the plane of its in-ring bonds. The first step generates all observed conformers for non-macrocyclic rings and the second step corrects for the inversions of stereochemistry in the first step (Figure 2.4 a). In addition, the ring-flipped version of all of these conformers are generated by reflecting the whole ring and then correcting the stereochemistry of all substituents as before (Figure 2.4b).

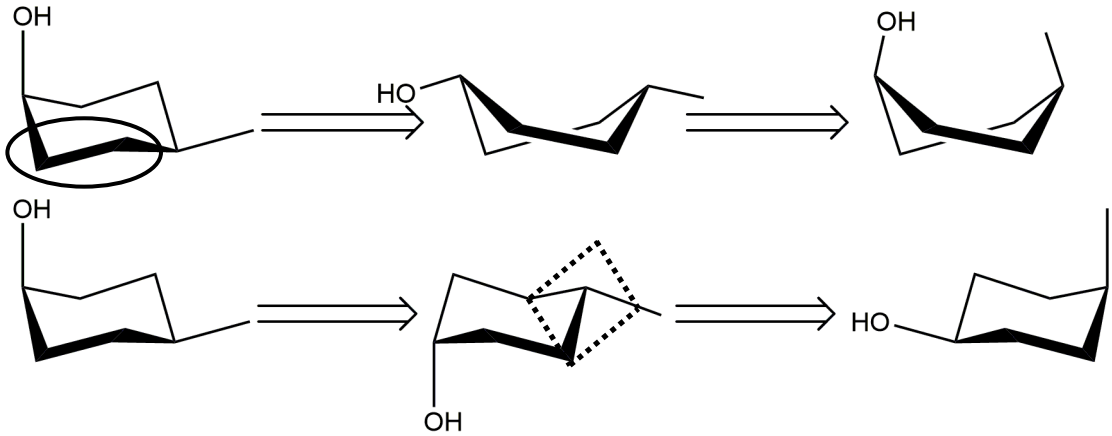



**Figure 2.4a)** The transformation of a chair conformer into a twist-boat conformer by a flip-of-fragments operation. Left: The chair conformer with the atoms being flipped circled. Middle: After the initial flip-of-fragments, the geometry of the ring itself is correct but all stereocenters are inverted. Right: After reflecting all substituents, the transformation is completed

**Figure 2.4b)** The interconversion of two chair conformers. Left: The chair conformer. Right: After reflecting the entire ring, the geometry of the ring is again correct but all stereocenters are inverted. This can be corrected by reflecting all substituents through the plane defined by the atom they are attached to and its in-ring meighbours (illustrated with dashed lines for the methyl group). Right: Once this second set of reflections is performed, the ring-flip is complete.

### 2.2.4: Screening for allowed conformers

Once a trial conformer has been generated, it is screened for excess steric crowding. If any two atoms separated by more than two bonds are closer than the scaled sum of their van der Waals radii, the conformer is rejected. Otherwise it is accepted and written to output. The sum of the van der Waals radii must be scaled so as to avoid rejection of conformers with favorable interatomic interactions such as hydrogen bonding. In such a case the atoms involved are by definition closer than the sum of their van der Waals radii. The van der Waals radii used are those of Mantina [17] and the scaling factor defaults to 0.7 as, in our experience, higher values may fail to generate conformer ensembles for extremely sterically crowded molecules.

## 2.3 Analysis

The UCONGA method can not only generate conformer ensembles, but also analyze them. There are three complementary analysis routines: filtering, clustering and visualization. These will now be discussed in greater detail.

### 2.3.1 Filtering

The filtering process serves two purposes. First, it removes redundant conformers that would be likely to optimize to the same local minimum if geometry optimization were performed. Second, it acts as a crude measure of the diversity of the conformer ensemble. This filtering is done using the root-mean-square deviation in atomic coordinates (RMSD) as a metric of the dissimilarity between two conformers. The conformers are compared in the order they are generated to all the already-accepted conformers. If any already-accepted conformer is too similar, defined as having a RMSD of 1.0 Å or less to the conformer under consideration, then it is rejected. To calculate the RMSD between two conformers, they are first aligned using the Schonemann [18] algorithm if enantiomeric conformers are designated as being identical or the Kabsch [19-20] algorithm if not. The RMSD is then calculated as follows:

$$d_{RMS} = \sqrt{\frac{1}{N_{atoms}} \sum_{i=1}^{N_{atoms}} \left| \boldsymbol{c}_{1,i} - \boldsymbol{c}_{2,i} \right|^2}$$

where $\boldsymbol{c}_{1,i}$ is the coordinate vector of the $i^{\text{th}}$ atom of conformer 1.

### 2.3.2 Clustering

The aim of clustering is to further reduce the dimensionality of the problem space, distilling out information on the main similarities and differences between the remaining conformers. In UCONGA, clustering is performed using the *k*-means algorithm. This groups conformers so as to minimize the within-cluster variance. The number of clusters can be determined using the Calinski-Harabasz criterion [21]. The Calinski-Harabasz criterion is the ratio of the between-cluster variance to the within-cluster variance, corrected for the number of clusters. It is therefore at a maximum for high-quality clustering where the clusters are tight and well-separated. There are two metrics used for clustering, one based on torsion angles and another based on the overall size of the

conformer. RMSD itself cannot be used as a metric for *k*-means clustering, as while it can be used to define a distance between two conformers, it cannot be used to calculate an average. By contrast, the average of torsion-angle vectors and bounding-box vectors can be calculated. The metric based on torsion angles, called the torsion-space distance, was calculated as

$$d_{\text{tors}} = \sqrt{\sum_{i=1}^{N_{\text{tors}}} \left( \left( \sin \varphi_{1,i} - \sin \varphi_{2,i} \right)^2 + \left( \cos \varphi_{1,i} - \cos \varphi_{2,i} \right)^2 \right)}$$

where $\varphi_{1,i}$ is the $i^{\text{th}}$ torsion angle of conformer 1. The trigonometric transformation of torsion angles used in the above formula was required to ensure this distance metric is Cartesian, as required by the *k*-means clustering algorithm.

The second metric was based upon the size of the molecule. More specifically, it is based on the dimensions of the bounding box, the smallest rectangular box that contains the molecule, after the coordinate axes are aligned to the three principle axes of rotation of the molecule. It is calculated as the Euclidian distance between the bounding box corners:

$$d_{\text{size}} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

where $(x_1,y_1,z_1)$ are the coordinates of the furthest corner of the smallest bounding box originating at (0,0,0) and completely containing conformer 1.

### 2.3.3 Visualization

During development and testing, visualization was important to aid understanding of the clustering results. There are three forms of visualization available. The first is a parallel coordinates plot of the clusters in torsion space, with the torsion identities on the *x* axis and the angles they adopt on the *y* axis. The second is a scatter plot of the two most important bounding box dimensions. In both of these, conformers can be color-coded by cluster identity. In addition, the RMSD matrix can be visualized using a heatmap where the similarity between conformers is represented by the color of elements within square matrix. The conformers can be sorted by their cluster identity for this visualization.

## 2.4 Implementation

The implementation of UCONGA is designed to make the program easy to use. It is written in python to avoid the need for compilation or installation. The only mandatory external dependency is NumPy [22], a high-performance linear algebra library that is common among scientific python users. Two other libraries that build on and are commonly installed with NumPy are required for access to all the features of the analysis module. The clustering functionality relies on SciPy [23], while the visualization relies on MatPlotLib [24]. In terms of design, three command-line programs are provided. All programs read the cml file format, which can be generated by the free software Avogadro [25] and OpenBabel [26] and all programs can write output as cml, xyz and the geometry portion of a GAMESS, Gaussian, or NWChem input file. This flexibility in output format generation allows for easier visualization or further optimization of the generated or aligned conformers. One program, called UCONGA_generate, generates the conformer ensemble. A second program, called UCONGA_analyse, performs clustering, RMSD calculation and visualization. The third program, called UCONGA_align, can align multiple conformers for viewing with a molecule viewer. The code is available at http://github.com/NRGunby/UCONGA under a 3-clause BSD license.

## 2.5 Benchmarking

### 2.5.1 The benchmark dataset

The UCONGA method was benchmarked using two sets of data, available in the electronic appendix. The first was the subset of the ASTEX dataset [27] containing five or fewer rotatable bonds. The ASTEX dataset contains high-quality crystal structures of ligands bound to proteins and is commonly used for benchmarking conformer-ensemble generation methods and therefore allows the comparison of UCONGA to other methods. The second dataset was a collection of molecular structures determined by gas-phase electron diffraction, including dimethylbis(trimethylsilylketyl)silane [28], tri-*tert*-butyl-sulfurtriimide [29], bis(*tert*-butyl)trichlorosilylphosphane [30], 1,1,2,2-tetra-*tert*-butyl-disilane [31] and 1,1,2-tri-*tert*-butyl-disilane [32]. These are molecules reported in the literature to have unusual conformers in the gas phase and are therefore a good test of the universal applicability of the UCONGA method. The size of the molecules in the combined dataset is summarized using a box-and-whisker plot in Figure 2.5.

In this and all other box-and-whisker plots, each vertical box spans half the data from the first to the third quartile with the median indicated within. Whiskers extend from the ends of the box to show the full data range, except for outliers further than 1.5 interquartile ranges from the median. These are represented by crosses.

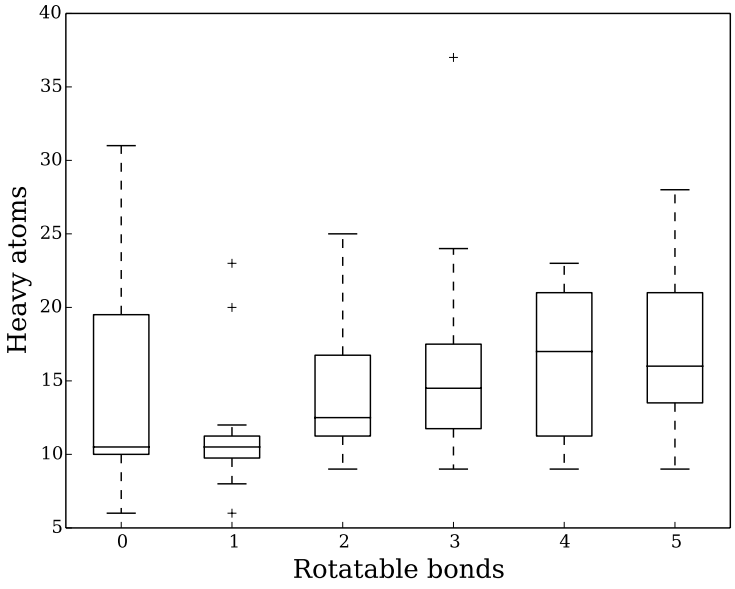



**Figure 2.5** A boxplot summarizing the size distribution of the molecules in the combination of the GED and utilized ASTEX datasets. The molecules contain enough rigid groups that their size is roughly independent of the number of rotatable bonds.

### 2.5.2 Algorithmic choices

The UCONGA method was used to generate conformer ensembles for all molecules in these datasets with 5 or fewer rotatable bonds. Two step sizes were used; an initial coarse step of 60° and a second fine step of 30°. No further geometry optimization was performed on the generated conformers. Before searching for the closest generated conformer to the reference, the reference was canonicalised so that atom labeling in symmetric rotors was consistent with what the UCONGA method would generate.

## 2.6 Results

In 78 of 84 cases, the generated ensemble included the experimentally determined conformer as determined by RMSD (Figure 2.6). In accordance with other conformer ensemble generation method benchmarking [4, 8], a non-hydrogen RMSD less than 1 Å between the best generated and experimentally determined reference conformers indicates a good fit, a heavy-atom RMSD between 1 and 2 Å indicates an acceptable fit and a greater heavy-atom RMSD indicates a poor fit.

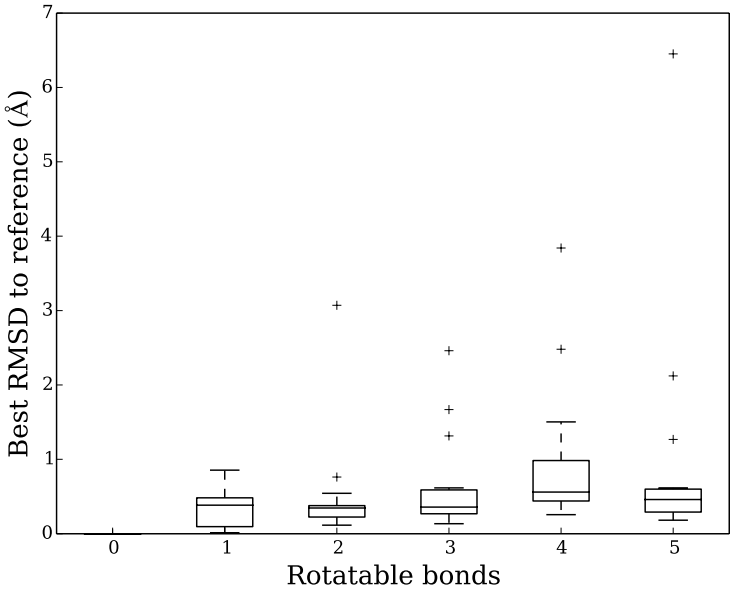



**Figure 2.6** In 93% of cases tests the conformer ensemble generated by UCONGA included the experimentally relevant conformer.

The size of the generated ensemble grows approximately exponentially with the size of the molecule (Figure 2.7). One notable exception is the collection of molecules with no rotatable bonds. These are solely due to ring conformers. Ring conformers are also responsible for the largest ensemble sizes.

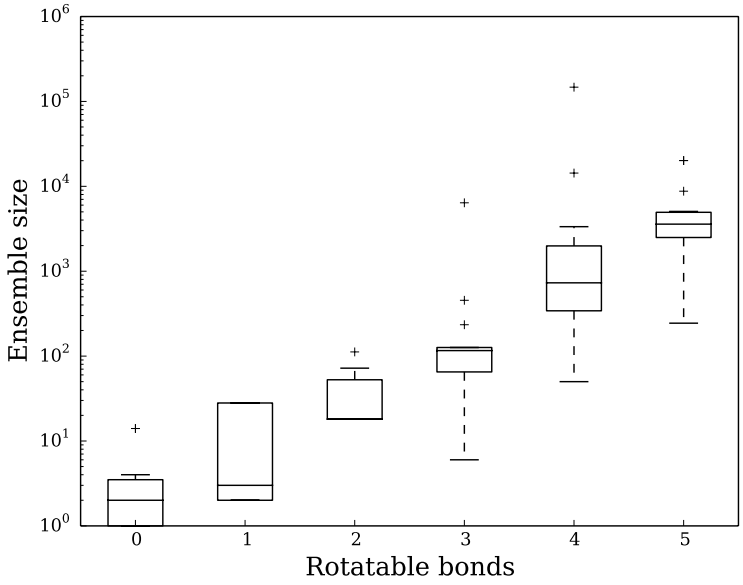



**Figure 2.7** The number of generated conformers grows rapidly with the number of rotatable bonds in the molecule.

When the conformer ensemble was filtered, the growth is still approximately exponential, but at a slower rate (Figure 2.8). The ratio of unfiltered:filtered ensemble sizes seems to converge to 10:1 for sufficiently large molecules, namely those with 4 or 5 rotatable bonds, although since there are only two post-convergence datasets this cannot be stated with certainty. This suggests that these conformer ensembles are highly redundant and not particularly diverse.

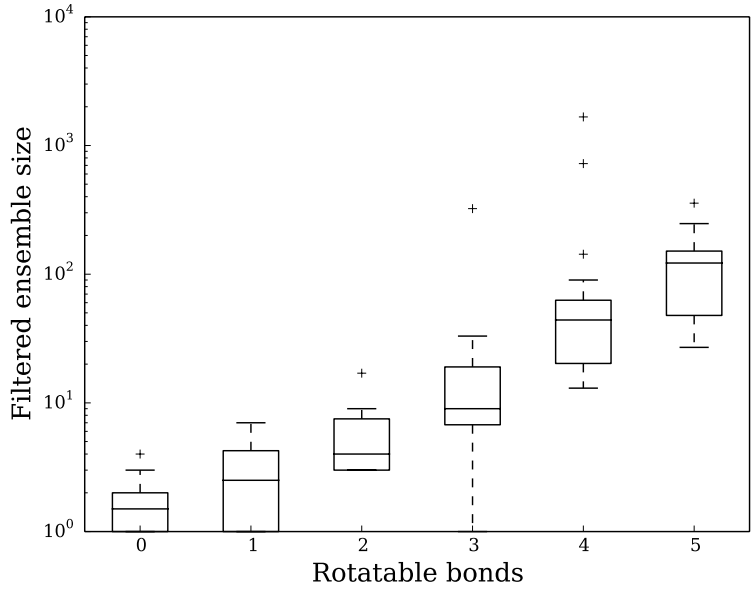



**Figure 2.8** The size of the filtered ensemble grows less rapidly than the unfiltered ensemble.

## 2.7 Discussion

It is informative to investigate the cases where UCONGA apparently failed to generate the reference conformer. One possibility is that RMSD may be a poor metric for assessing structural similarity as it may be artificially inflated by small misrotations of rigid and/or bulky terminal groups. For example, a chain of rotatable bonds capped by a 9-anthracenyl group will have a higher RMSD difference for a given difference in torsion angles than the same chain capped by a phenyl group, because the anthracenyl group has a higher average radius between its carbon atoms and the axis of rotation (Figure 2.9).

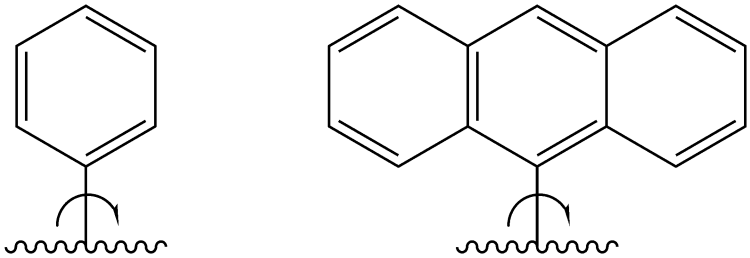


**Figure 2.9** The RMSD difference caused by rotation around a bond is proportional to the radius between the atom and the bond, which is larger for a 9-anthracenyl group (right) than a phenyl group (left).

A useful way of checking this is to compare the conformers chosen as the closest to the reference using both RMSD and torsion-space distance. This is done in Figure 2.10, sorting molecules according to whether the two metrics agree on the closest conformer to the reference, whether there is near-agreement, defined as at least one conformer in common between the two lists of the three most similar conformer to the reference, or whether there is poor agreement, defined as all other cases.

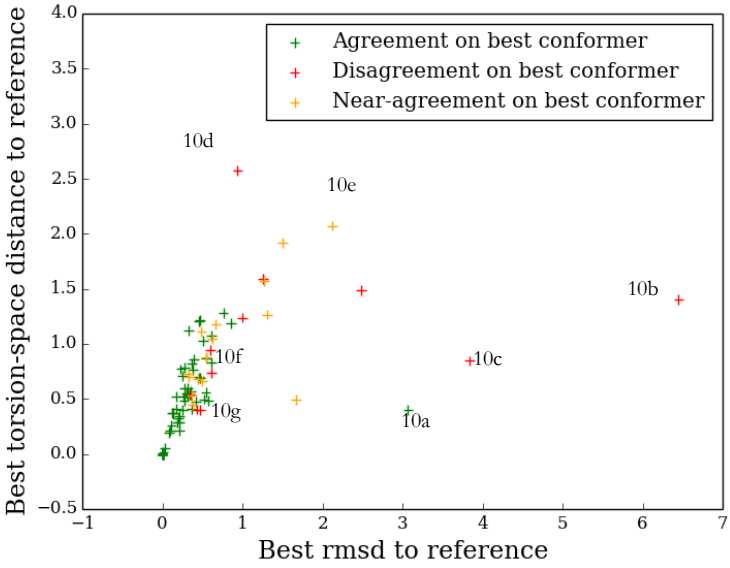



**Figure 2.10** A comparison of the quality of the best conformer by the two metrics, with points of interest labeled. Their structures are given in Figure 2.11.

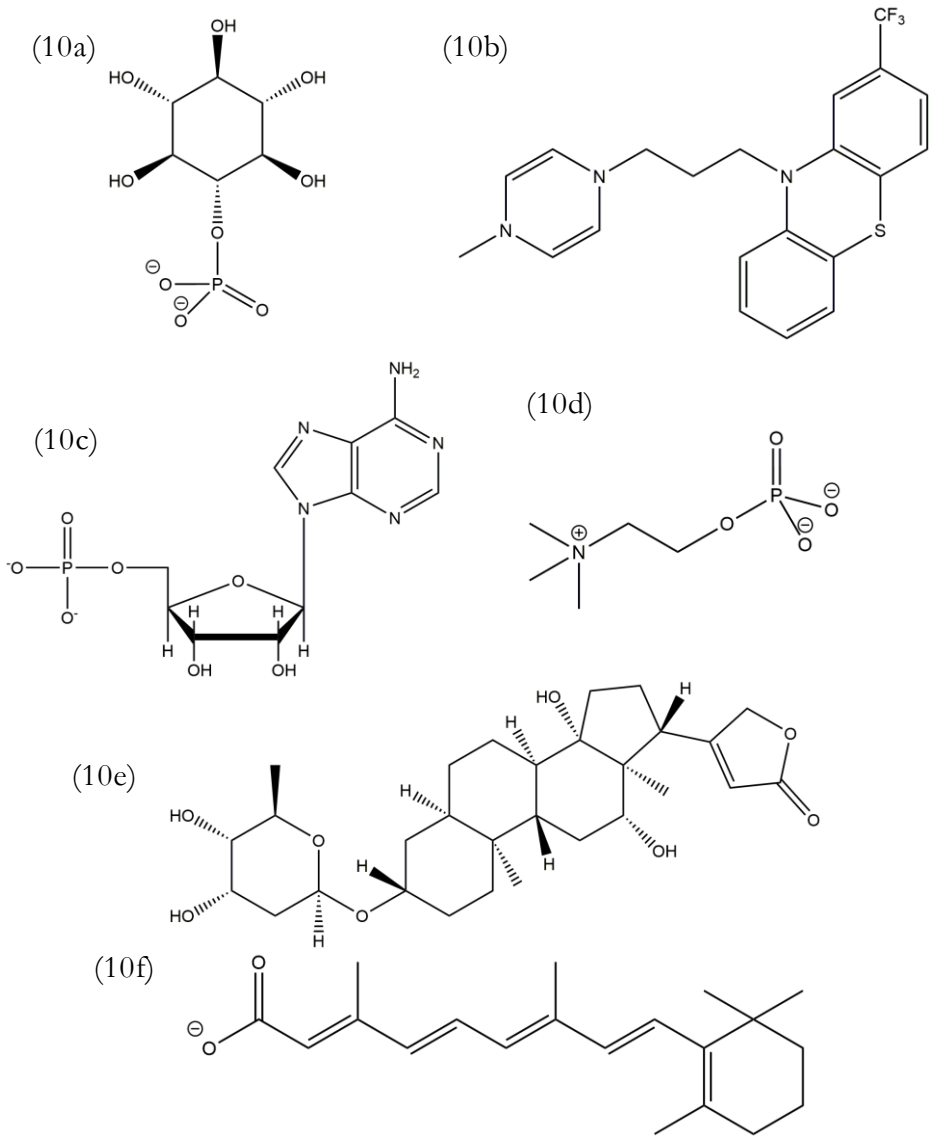



**Figure 2.11** The structures of the outliers labeled on Figure 2.10

There are four types of outliers. For 10a, both metrics agree on the closest conformer to the reference, but disagree on how close it is. For 10b-d, the metrics disagree on which conformer is closest and only one of them identifies one that is close. For 10e, the metrics are in near-agreement as to the closest conformer to the reference, but neither method identifies a conformer that is particularly close. For 10f and 10g, the metrics disagree as to what the closest conformer is, but both have found one that is close.

### 2.7.1 Molecule 10a

One outlier, marked as 10a on Figure 2.10, has the same conformer identified as the closest to the reference. It is predicted to be a good fit based on torsion-space distance but a poor fit by RMSD. The molecule is myo-inositol-1-phosphate. The superposition of the reference and best conformers (Figure 2.12) suggests that the best conformer is an

acceptable fit: the conformer of cyclohexane is correct and three of the oxygen atoms of the phosphate group are close. On further investigation, it seems that this disagreement is due to a failure of the canonicalization: the near-overlapping pairs of oxygen atoms have different labels despite the fact that the canonicalization process should ensure the same atom labeling sequence in the reference and UCONGA-generated conformers. This is a problem only when comparing conformers generated by UCONGA with conformers generated by other methods; when UCONGA is the only source of conformers canonicalization is irrelevant.

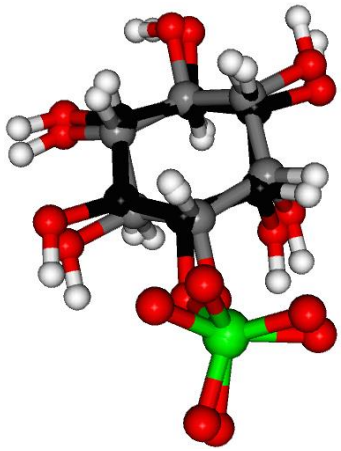


**Figure 2.12** The reference (black carbon skeleton) and best generated (grey carbon skeleton) conformers of myo-inositol-1-phosphate.

### 2.7.2 Molecules 10b-d

The worst fit by RMSD is for 10b, 10-[3-(4-methylpiperazin-1-yl)propyl]-2-(trifluoromethyl)-10*H*-phenothiazine, better known as the antipsychotic drug trifluoperazine. As shown in Figure 2.13, the RMSD alignment heavily weights the phenothiazine ring system, which is most relevant for the shape of the molecule in the computational ligand-protein docking applications where RMSD is heavily used. Torsion-space distance, on the other hand, choses a conformer that is a very close match except for the phenothiazine ring system. The difference between the selections is due to the propensity of the RMSD metric to over-weight bulky terminal groups, as shown in Figure 2.12.

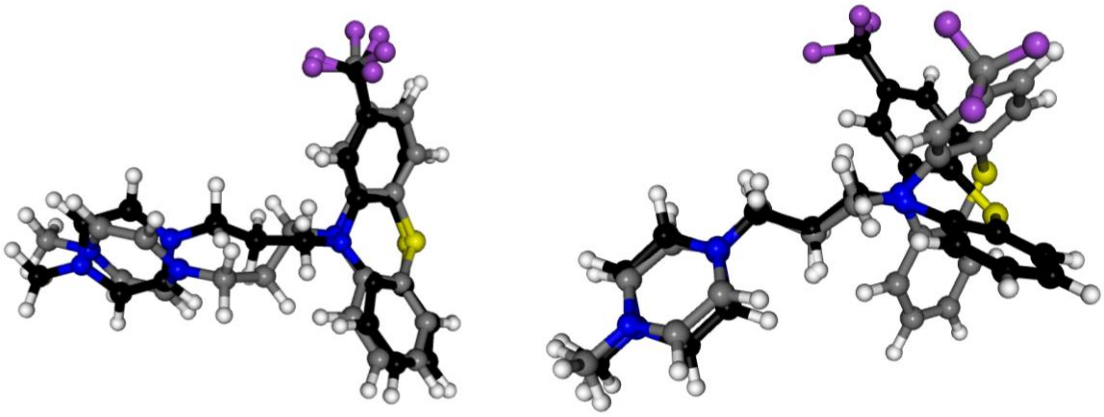


**Figure 2.13** The reference (black carbon skeleton) and best generated (grey carbon skeleton) conformers of trifluoperazine according to RMSD (left) and torsion-space distance (right).

The situation is similar for 10c, adenosine monophosphate (Figure 2.14), which has with the second-worst fit by RMSD. In the conformer chosen by RMSD, the large terminal ring is closer to the reference, but the backbone (specifically the methoxy group) is closer to the reference in the conformer chosen by the torsion-space distance.

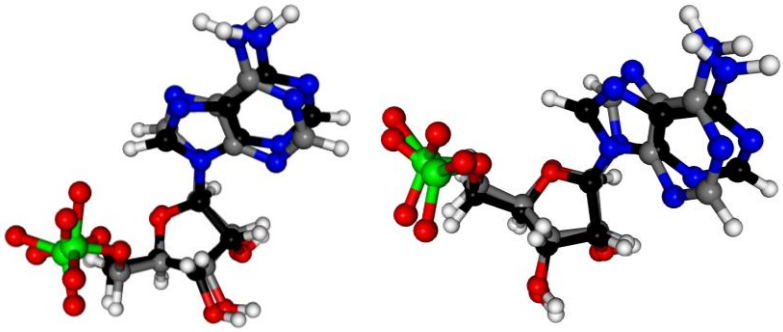


**Figure 2.14** The reference (black carbon skeleton) and best generated (grey carbon skeleton) conformers of adenosine monophosphate according to RMSD (left) and torsion-space distance (right).

The one outlier with better RMSD than torsion-space distance is 10d, phosphatidylcholine (Figure 2.15). Analogously to trifluoperazine, in the conformer chosen and aligned by RMSD, the bulkiest end of the molecule overlaps reasonable well, with the rest of the molecule having a similar radius from the axes between the two ends but taking a different path along the surface of the cylinder. Torsion-space distance once again aligns the backbone better but the termini worse. The problem again seems to be incorrect canonicalization of the phosphate group. As mentioned above, this is only a problem when comparing conformers generated by UCONGA with conformers generated by other methods and is irrelevant for normal usage of UCONGA.

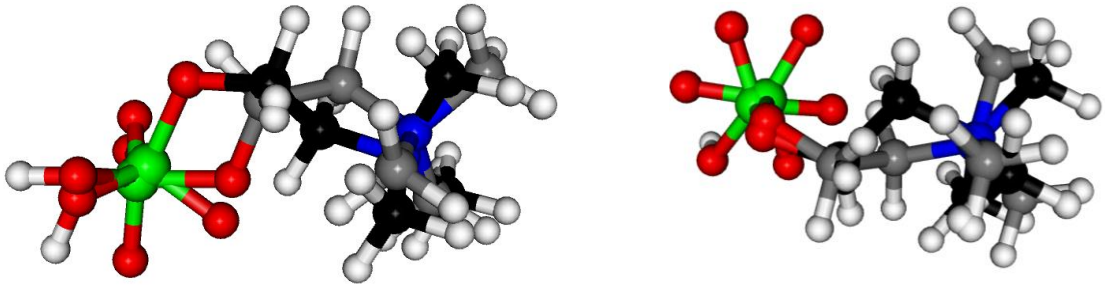


**Figure 2.15** The reference (black carbon skeleton) and best generated (grey carbon skeleton) conformers of choline according to RMSD (left) and torsion-space distance (right).

### 2.7.3 Molecule 10e

A straightforward explanation for poor fit can be found for 10e, 1,1,2,2-tetra-*tert*-butyl-disilane, a molecule from the GED test set where there is near-agreement on the best conformer by the two metrics and where neither of the best conformers fits particularly well. This is because the reference conformer itself would be rejected by UCONGA, even with the 0.7 scaling factor for the van der Waals radius. Investigating further, this is because the bond angles and torsion angles are coupled, with the experimental structure featuring an Si-Si-C bond angle of almost 120° to the eclipsed *tert*-butyl groups which UCONGA is incapable of reproducing as it only alters torsion angles and ring conformers. This should not be a general problem as this degree of distortion indicates that the limit of steric crowding that is possible around a bond is being reached, hence molecules with a greater degree of steric crowding are unlikely to be stable.

### 2.7.4 Molecules 10f-g

Finally, for 10f and 10g, both metrics identify closely matching conformers to the reference, but disagree on the identity of the closest. In both cases (Figure 2.16), the conformer chosen by torsion-space distance is visually a better fit than that identified by RMSD.

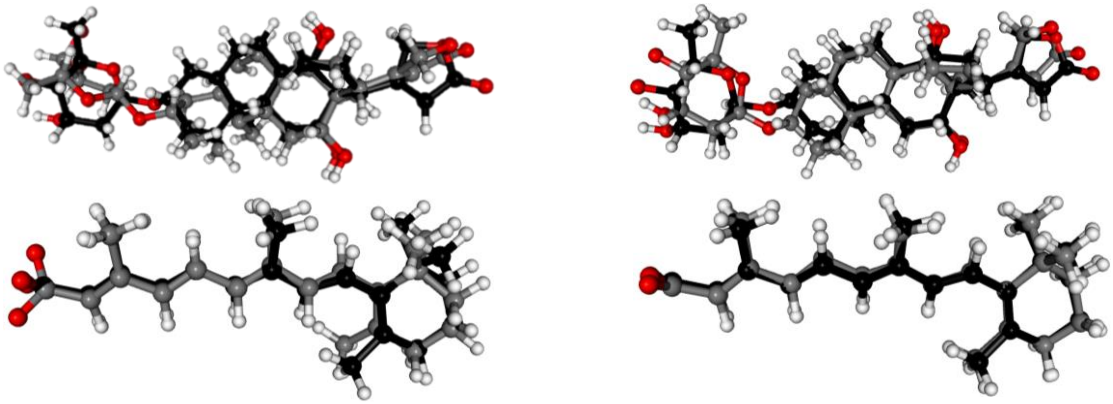


**Figure 2.16** The reference (black carbon skeleton) and best generated (grey carbon skeleton) conformers of 10f (top) and 10g (bottom) according to RMSD (left) and torsion-space distance (right).

### 2.7.5 Summary

Summarizing these cases, three things can be noted. Firstly, RMSD as a metric is more sensitive to small changes in a large terminal group than to relatively large changes in an unbranched chain of rotatable bonds. This is desirable behavior for computational docking studies with no optimization between conformer generation and docking, but may be undesirable for other applications. Secondly, that when the two metrics disagree on which conformer is closest to the reference, that identified by torsion-space distance is typically a better fit and more likely to optimize to the target structure if a subsequent geometry optimization step is performed. Third, many of the phosphate groups are poorly canonicalized, due to difficulty in doing this for these groups. If the torsion in the reference conformer is close to one of the end-points of the unique torsion values of this rotamer, the version of the reference that UCONGA would generate is not necessarily the best fit to the closest conformer that UCONGA actually generates, as shown in Figure 2.17. These canonicalization difficulties are only relevant when comparing conformers generated using UCONGA to those generated using some other method. The oxygen atom labeling in phosphate groups is irrelevant for normal use of UCONGA.

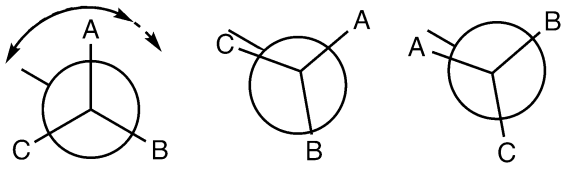



**Figure 2.17** For a triply symmetrical rotor like a phosphonate group, each oxygen atom (here labeled A, B, and C) has 119° of unique rotation (left, dashed arrow) but, assuming a 30° step size, will only have conformers generated in 90° of that space (left, solid arrow). A torsion angle of 110° is canonical (middle), but each oxygen atom is 20° away from its position in the closest generated conformer. If the atom labeling was non-canonical (right), with a torsion angle of -10°, each oxygen atom would only be 10° away from a position that could be generated.

## 2.8 Conclusion

The newly developed UCONGA method is capable of generating conformer ensembles that almost always contain the experimentally relevant conformer for molecules as diverse as amino acids and sterically hindered asymmetric disilanes. UCONGA is fully general with respect to the structure of the molecule and seems able to generate diverse conformers. However, it is somewhat limited in the sizes of the molecules it can generate conformer ensembles for, as generating conformers for molecules with more than five rotatable bonds would take an unfeasible amount of time. Additionally, the exponential increase of ensemble size with the number of rotatable bonds makes using the ensembles for larger molecules challenging. Filtering the ensemble using RMSD can help decrease the ensemble size, but not the time taken. Some modifications to the UCONGA method are required to handle these challenges. These modifications will be discussed further in Chapter 3.

## 2.9 References


1) Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discov. Today Technol.* **2010**, *7*, e245–e253.
2) Klett, J.; Cortés-Cabrera, Á.; Gil-Redondo, R.; Gago, F.; Morreale, A. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.* **2014**, *54*, 314–323.
3) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.
4) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.
5) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.
6) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.
7) Mitzel, N. W.; Rankin, D. W. H. SARACEN – Molecular Structures from Theory and Experiment: The Best of Both Worlds. *Dalton Trans.* **2003**, (19), 3650–3662.
8) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.
9) Donadel, O. J.; Martín, T.; Martín, V. S.; Villar, J.; Padrón, J. M. The *tert*-Butyl Dimethyl Silyl Group as an Enhancer of Drug Cytotoxicity against Human Tumor Cells. *Bioorganic Med. Chem. Lett.* **2005**, *15*, 3536–3539.
10) Franz, A. K.; Wilson, S. O. Organosilicon Molecules with Medicinal Applications. *J. Med. Chem.* **2013**, *56*, 388–405.
11) Bikzhanova, G.; Toulokhonova, I.; Gately, S.; West, R. Novel Silicon-Containing Drugs Derived from the Indomethacin Scaffold: Synthesis, Characterization and Evaluation of Biological Activity. *Silicon Chem.* **2007**, *3*, 209–217.
12) Van Hattum, A. H.; Pinedo, H. M.; Schlüper, H. M. M.; Hausheer, F. H.; Boven, E. New Highly Lipophilic Camptothecin BNP1350 is an Effective Drug in Experimental Human Cancer. *Int. J. Cancer* **2000**, *88*, 260–266.

13) Lloyd, N. C.; Morgan, H. W.; Nicholson, B. K.; Ronimus, R. S. The Composition of Ehrlich's Salvarsan: Resolution of a Century-Old Debate. *Angew. Chem. Int. Ed.* **2005**, *44*, 941–944.

14) Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry. *J. Chem. Inf. Model.* **1977**, *17*, 110–113.

15) Razinger, M.; Balasubramanian, K.; Perdih, M.; Munk, M. E. Stereoisomer Generation in Computer-Enhanced Structure Elucidation. *J. Chem. Inf. Model.* **1993**, *33*, 812–825.

16) Mekenyan, O.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45*, 283–292.

17) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. Consistent van der Waals Radii for the Whole Main Group. *J. Phys. Chem. A* **2009**, *113*, 5806–5812.

18) Schönemann, P. H. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika* **1966**, *31*, 1–10.

19) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923.

20) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

21) Caliński, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27.

22) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

23) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python http://www.scipy.org/ (accessed Mar 26, 2015).

24) Hunter, J. D. Matplotlib: A 2D Graphic Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

25) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminform.* **2012**, *4* (1), 17.

26) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.

27) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins Struct. Funct. Genet.* **2002**, *49*, 457–471.

28) Borisenko, K. B.; Yezhov, R. N.; Gruener, S. V.; Robertson, H. E.; Rankin, D. W. H. Gas-Phase Molecular Structures of Substituted 1, 3-Bisketenes: A Challenge for Theory and Experiment. *Inorg. Chim. Acta* **2008**, *361*, 467–472.

29) Hinchley, S. L.; Trickey, P.; Robertson, H. E.; Smart, B. A.; Rankin, D. W. H.; Leusser, D.; Walfort, B.; Stalke, D.; Bühl, M.; Obrey, S. J. Bis(*tert*-butyl)sulfurdiimide, $S(NBu^t)_2$, and Tris(*tert*-butyl)sulfurtriimide, $S(NBu^t)_3$: Structures by Gas Electron Diffraction, X-Ray Crystallography and *Ab Initio* Calculations. *J. Chem. Soc., Dalton Trans.* **2002**, 4607–4616.

30) du Mont, W. W.; Müller, L.; Martens, R.; Papathomas, P. M.; Smart, B. A.; Robertson, H. E.; Rankin, D. W. H. Intermediates and Products of the Hexachlorodisilane Cleavage of Group 14 Element Phosphanes and Amines – Molecular Structure of Di-*tert*-butyl(trichlorosilyl)phosphane in the Gas Phase Determined by Electron Diffraction and *Ab Initio* Calculations. *Eur. J. Inorg. Chem.* **1999**, 1381–1392.

31) Hinchley, S. L.; Robertson, H. E.; Parkin, A.; Rankin, D. W. H.; Tekautz, G.; Hassler, K. Molecular Structure of 1,1,2,2-Tetra-*tert*-butyldisilane: Unusual Structural Motifs in Sterically Crowded Disilanes. *Dalton Trans.* **2004**, (5), 759–766.

32) Hinchley, S. L.; Smart, B. A.; Morrison, C. A.; Robertson, H. E.; Rankin, D. W. H.; Zink, R.; Hassler, K. 1, 1, 2-Tri-*tert*-butyldisilane, $Bu^t_2HSiSiH_2Bu^t$: Vibrational Spectra and Molecular Structure in the Gas Phase by Electron Diffraction and *Ab Initio* Calculations. *J. Chem. Soc., Dalton Trans.* **1999**, 2303–2310.

# Chapter 3

## Development of a divide-and-conquer method for UCONGA

## 3.1 Introduction

As discussed in Chapter 2, while the UCONGA method is universal with respect to the structural features of a molecule, it is limited with respect to the size of the molecules it can practically be applied to. The UCONGA method can typically only generate conformer ensembles for molecules with 5 or fewer rotatable bonds, except for highly symmetrical molecules with a high degree of steric crowding. Considering that more specialized conformer ensemble generation methods can generate conformer ensembles for molecules with 12 or more rotatable bonds, as demonstrated in a recent comparison [1], this is a significant limitation.

Divide-and-conquer algorithms are a class of method that often have good performance with respect to the size of the problem. As the name suggests, they involve breaking a complex problem into a series of subproblems that are solved directly and independently, and then recombined to approximate the global solution. If each sub-problem is truly independent then the exact global solution will be recovered. In the context of conformer generation, divide-and-conquer approaches work by fragmenting the molecule and then finding fragment conformers. Fragment conformers are then systematically recombined in all possible combinations. This reduces the computational cost and complexity of generating the conformer ensemble if many fragment conformers are rejected before the recombination step, as this reduces the overall number of trial conformers tested compared with direct conformer generation.

Divide-and-conquer methods have been used by other methods for conformer generation. The OMEGA method [2] splits a molecule into rings and linear linkers, whose conformations it reads from a database, and rotates about the bonds between them. The Confgen method [3] performs a potential energy surface search about each isolated bond and then finds conformers by combining minima from those. It is therefore reasonable to assume that a divide-and-conquer algorithm in which the fragment conformers are generated using the UCONGA algorithm would be able to generate conformer ensembles for larger molecules more efficiently than UCONGA alone.

## 3.2 Algorithm

The divide-and-conquer algorithm for the UCONGA method fragments a molecule, finds conformer ensembles for each fragment and then systematically recombines all possible combinations of fragment conformers. Each of these recombined conformers is screened for steric clashes as described in Chapter 2. For this to improve efficiency, as many fragment conformers as possible should be eliminated at the fragment stage, therefore also reducing the number of recombined conformers. It is therefore advantageous to fragment the molecule so each fragment contains bonds that adopt coupled torsion angles. Typically, this involves fragmenting the molecule at rigid linkers, and including each rigid linker unit in each generated fragment.

The most general way to achieve this is to ensure all adjacent rotatable bonds are included within fragments wherever possible. For molecules with more than five rotatable bonds, this level of fragmentation will not be sufficient, as the number of fragment conformers will be too large to be practical. In other words, the conformational search problem for each fragment alone becomes prohibitive, to say nothing of recombination. Therefore, fragments containing more than five rotatable bonds will be subdivided as evenly as possible. This fragmentation pattern is similar to that used by OMEGA, but here the fragment conformers are not read from a database, as they are in OMEGA, rather they are generated using the UCONGA algorithm. Rigid linking groups are included in both fragments to increase the steric hindrance in each fragment and thus hopefully increase the number of rejected fragment conformers. Figure 3.1 illustrates the fragmentation process for a general molecule containing cyclic and multiply-bonded acyclic rigid linkers separating groups of connected rotatable bonds.

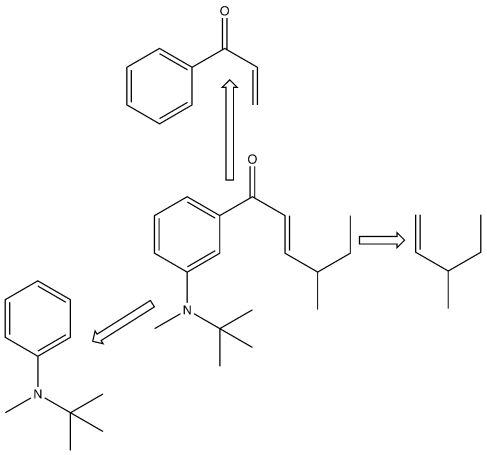



**Figure 3.1** A molecule and the fragments produced from it, each derived from a pair of adjacent rotatable bonds.

The black-box fragmentation algorithm used by UCONGA may not always be the most chemically sensible or efficient. For example, UCONGA would fragment conjugated polyene chains that should remain intact according to chemical intuition. Therefore, the program allows the user to select bonds to be broken to form fragments instead of accepting the default fragmentation. Flexible rings are held rigid for fragment conformer generation. Ring conformers can optionally be generated at the end of the recombination stage, using either the full flip-of-fragments algorithm discussed in Section 2.2.3 or a restricted method that only generates the input ring conformer and its reflection (see Figure 3.2 for an illustration of the results of these choices). Ring conformer generation is optional because ring conformers can cause a large increase in conformer ensemble size, which for some molecules may cause impractically high runtime or memory use. The conformer generation algorithm is summarized in Figure 3.3.

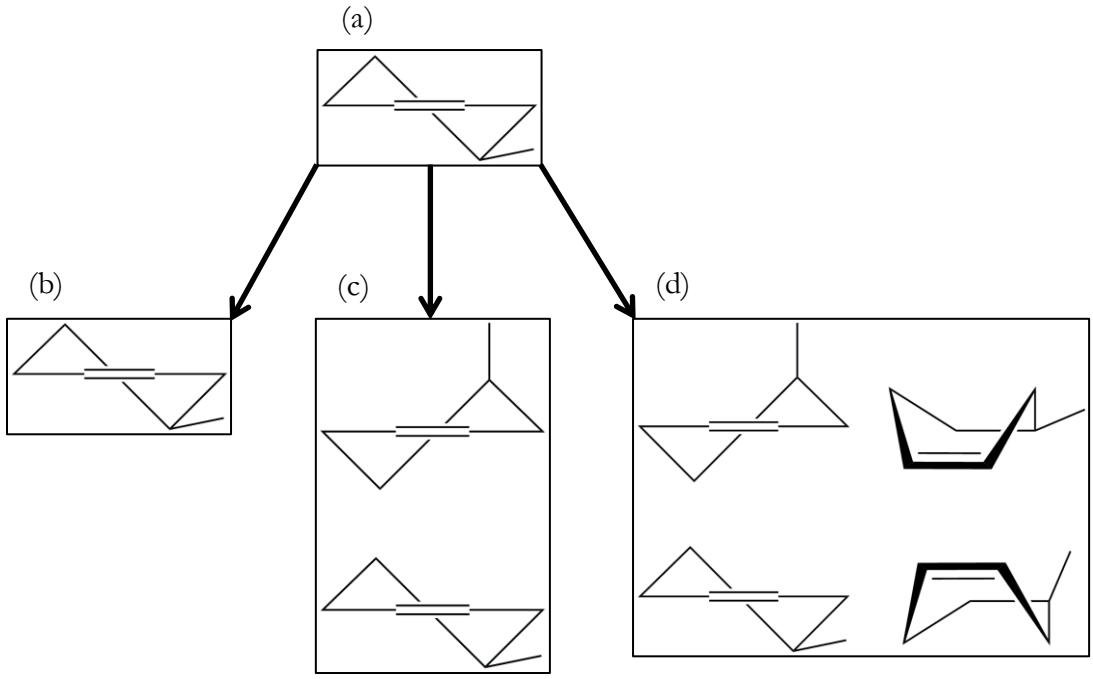



**Figure 3.2** The three options for ring conformer generation with the divide-and-conquer algorithm, illustrated using 4-methylcyclohexene: (a) A single half-chair conformer of cyclohexene, before ring conformer generation occurs; (b) If the user chooses to hold ring conformers constant, that same half-chair conformer is all that is generated; (c) If the user chooses restricted ring conformer generation, with reflection only, then both half-chair conformers are generated; (d) If the user chooses unrestricted ring conformer generation, both half-chair and both boat conformers are generated.

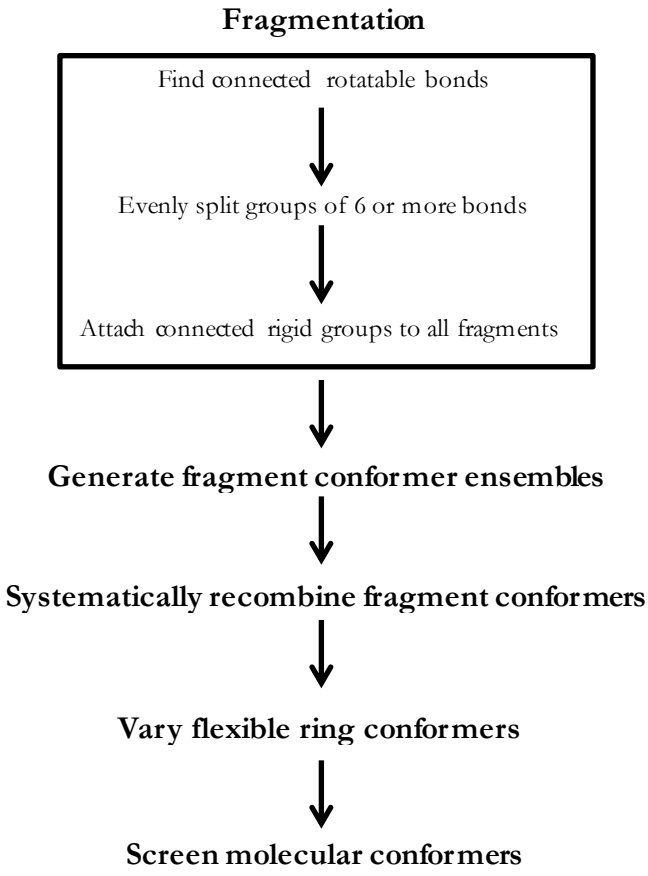



**Figure 3.3** A flowchart showing the divide-and-conquer algorithm of UCONGA

## 3.3 Methods

### 3.3.1 Standard UCONGA with restricted ring conformers

The divide-and-conquer algorithm incorporates optional ring conformer restriction, as outlined in Figure 3.2. For practicality, the reflection-only option, also called restricted ring conformer generation, will be used throughout when benchmarking the divide-and-conquer algorithm. For comparability, this same process of restricting ring conformer generation was applied to the original standard UCONGA algorithm described in Chapter 2. All algorithmic choices were otherwise kept constant.

To quantify ensemble diversity loss due to ring conformer restriction, ensembles generated using standard UCONGA both with and without ring conformer restriction were compared using clustering-reclustering as described in Figure 3.4. The diversity of the conformer ensembles generated with restricted and unrestricted rings were compared using bounding-box-based clustering. Cluster centers were found for the conformer ensembles generated with the divide-and-conquer algorithm. The distance from each

conformer to the closest cluster center was measured. The standard UCONGA conformers were then clustered using these cluster centers. Two metrics were then calculated for each cluster in both the standard and divide-and-conquer conformer ensembles. The first is the fraction of the ensemble in each of the clusters. The second metric is maximum distance from a conformer in each cluster to the center of that cluster.

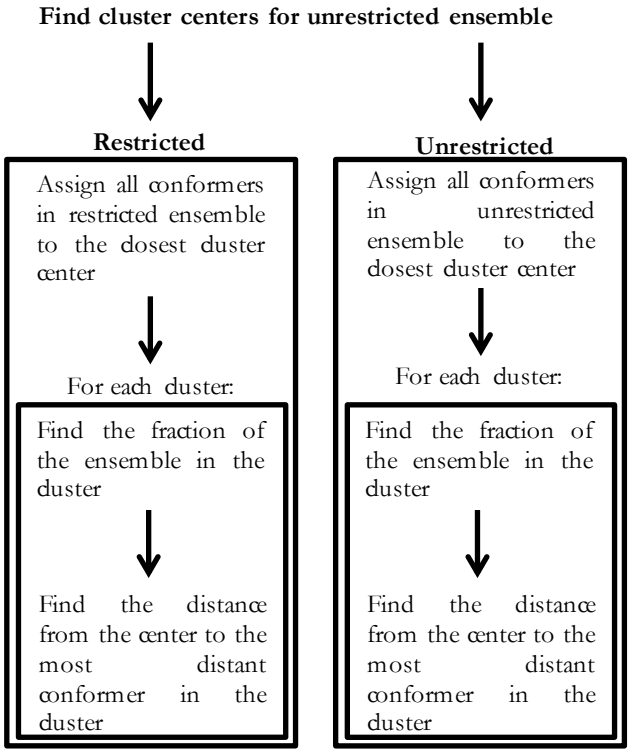



**Figure 3.4** A flowchart explaining how the diversity of the restricted and unrestricted conformer ensembles were compared.

### 3.3.2 Divide-and-conquer

To compare the effects of the divide-and-conquer and standard UCONGA algorithms, the divide-and-conquer algorithm was benchmarked on molecules from the ASTEX dataset with 4 or 5 rotatable bonds that would generate multiple fragments using the division algorithm described above. To test its scaling with system size, it was then used to generate conformer ensembles for molecules from the ASTEX data set with 6-8 rotatable bonds. These structures are given in the electronic appendix. Algorithmic choices are the same as for Section 3.3.1. As for benchmarking standard UCONGA, both RMSD and torsion-space distance were used to identify the closest generated conformer to the reference, and the percentage of the conformer ensemble remaining after it was filtered using RMSD was used as a crude measure of the diversity of the ensemble.

One change was made with the analysis methods; given the problems with canonicalization of phosphate groups discussed previously in Section 2.7 all permutations of oxygen atom labels were attempted and the one giving the best fit was used. The distribution of molecular sizes in the benchmarking set is shown in Figure 3.5.

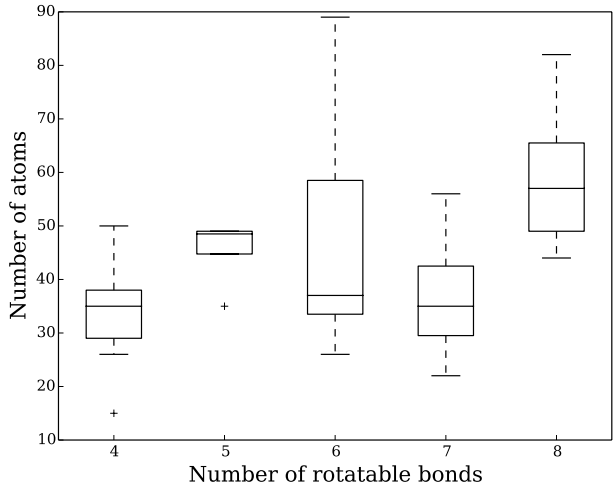



**Figure 3.5** The distribution of molecular sizes as a function of the number of rotatable bonds within each molecule in the benchmarking data set.

## 3.4 Results and discussion

### 3.4.1 Ring flexibility restrictions

Restricting the flexibility of rings reduced most ensemble sizes twofold to tenfold (Figure 3.6). The three smallest ensembles were not affected at all, and the largest ensemble was reduced by a factor of 181 from nearly 147,000 conformers to 810.

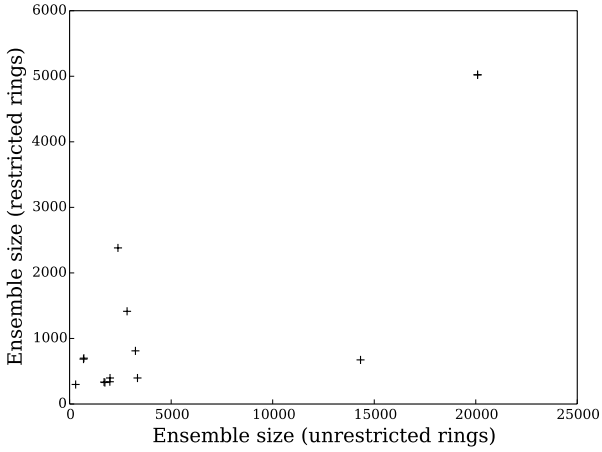



**Figure 3.6** A scatterplot comparing the ensemble sizes of molecules studied using standard UCONGA and both methods of generating ring conformers. There is an outlier not shown with an unrestricted ensemble size of nearly 147,000 conformers and a restricted ensemble size of 810 conformers.

This is accompanied by a loss of conformer diversity, as measured using bounding-box clustering as discussed above. The proportion of the conformer ensemble in each bounding-box cluster changes, and the conformers that are not generated sometimes include some of those furthest from the cluster centers. This is shown in Table 3.1 Occasionally, one of the clusters gets slightly bigger, which is due to numerical errors when aligning molecules to their axes of inertia (see Figure 3.7 as an example). Based on an examination of these, until a better alignment algorithm is implemented which is less prone to propagating numerical error, bounding-box measurements should be treated as having an accuracy of 0.3 Å.

**Table 3.1** The results of bounding-box-based clustering for all ensembles studied using standard UCONGA both in Chapter 2 with unrestricted ring conformer generation and in this chapter with restricted ring conformer generation.[a]

| ASTEX ID code | Cluster sizes (%) | | Cluster radius (Å) | |
|---|---|---|---|---|
| | Unrestricted | Restricted | Unrestricted | Restricted |
| 1aco | 52/47 | 52/47 | 2.33/3.01 | 2.33/3.01 |
| 1ase | 64/35 | 61/38 | 4.04/4.43 | 3.77/4.74 |
| 1cbs | 67/32 | 58/41 | 6.00/8.69 | 5.72/8.07 |
| 1cdg | 33/66 | 31/68 | 4.42/4.26 | 4.04/4.06 |
| 1ctr | 55/44 | 53/46 | 6.09/7.18 | 6.09/6.59 |
| 1epb | 33/66 | 51/48 | 8.82/7.87 | 8.64/7.87 |
| 1eta | 34/65 | 38/61 | 7.39/6.70 | 7.46/6.67 |
| 1fen | 40/59 | 38/61 | 7.26/7.08 | 6.10/5.27 |
| 1rob | 58/41 | 69/30 | 4.17/4.69 | 4.48/4.60 |
| 1tdb | 33/66 | 37/62 | 6.84/4.80 | 5.95/4.38 |
| 1tpp | 36/63 | 36/63 | 5.02/4.42 | 5.02/4.42 |
| 1ukz | 38/61 | 53/46 | 6.37/4.36 | 4.65/4.36 |
| 2ak3 | 57/42 | 39/60 | 3.72/6.89 | 3.53/5.43 |
| 2yhx | 45/54 | 45/54 | 6.86/4.66 | 6.44/4.90 |
| 6rnt | 64/35 | 62/37 | 5.04/5.24 | 4.19/3.88 |

[a]The Calinski-Harabasz criterion found two clusters for all systems, and the slash separates the values for the two clusters. The cluster size refers to the proportion of conformers in the two clusters, and the radius of the cluster refers to the distance from the cluster center to the furthest conformer in the cluster for the two clusters, using the distance metric for bounding-box clustering given in Section 2.3.2.

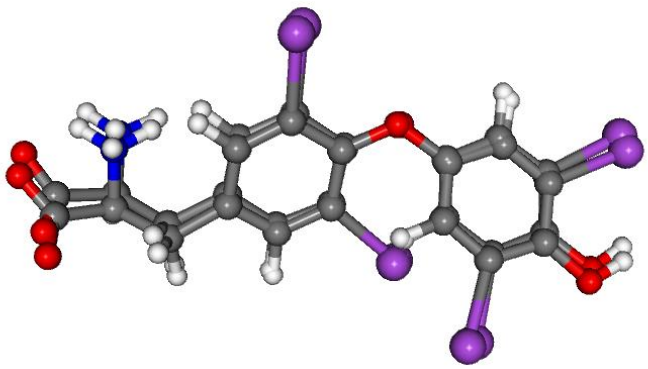


**Figure 3.7** The structures of the conformers of molecule 1eta from the restricted and unrestricted ensembles furthest from the center of the second cluster. While these are the same conformer, they are not perfectly aligned, which is especially noticeable at the carboxylate group, and thus have slightly different bounding boxes.

For molecules with six-membered rings the reduction in ensemble size is worth the loss of diversity as the ring conformers that are only generated with unrestricted ring conformer generation are boat conformers. For most six-membered rings these are higher in energy than the chair or half-chair conformers that are generated with restricted ring conformer generation. For molecules with five-membered rings, the situation is more complicated. There are five pairs of enantiomeric envelope conformers, each with a different atom out of the plane of the other four. Only the unrestricted ring conformer generation method can generate all pairs.

### 3.4.2 Comparing divide-and-conquer with standard UCONGA

For molecules that had previously been studied with standard UCONGA, the best RMSD to the reference conformer was similar to that found with standard UCONGA, so using the divide-and-conquer algorithm does not result in an inability to reproduce the reference conformer, at least for the molecules in the test set (Figure 3.8). In addition, the ensemble size is identical to that produced using standard UCONGA with the ring size restrictions (Figure 3.9).

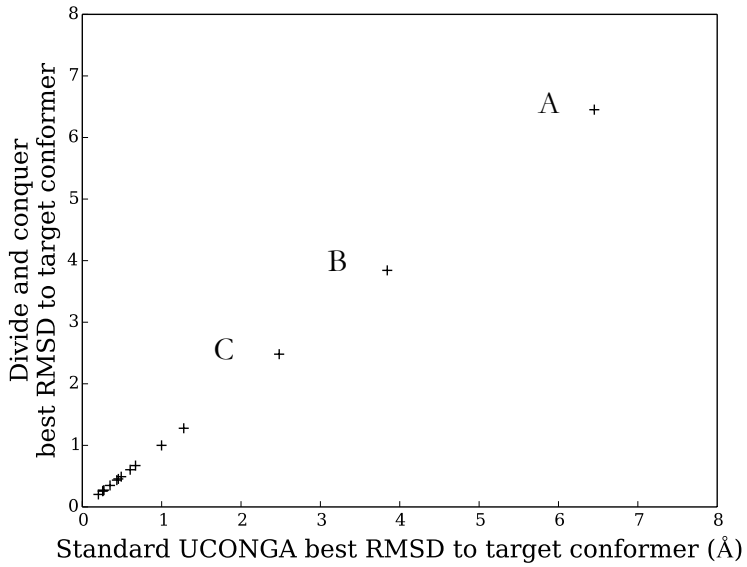



**Figure 3.8** A comparison of the RMSDs of standard and divide-and-conquer UCONGA-generated conformers that best reproduce the reference conformer. Outliers labelled A and B were discussed in Chapter 2, while the outlier labelled C will be discussed later in this chapter.

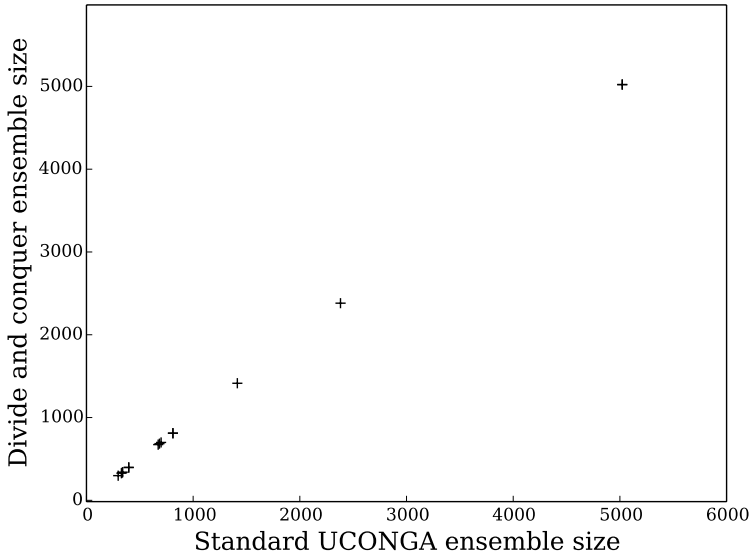



**Figure 3.9** A comparison of the sizes of standard and divide-and-conquer UCONGA-generated conformer ensembles.

This warrants further investigation, as there is a good reason to expect the standard UCONGA conformer ensemble to be larger due to the two-step trial conformer generation. This is because the second, smaller step (30°) is only taken around conformers generated with the first, larger (60°) step, and all torsions are rotated using this second step. In standard UCONGA, all torsions in the molecule will be rotated in the smaller step but in the divide-and-conquer algorithm, only torsions in the fragment that is rejected will be rotated with the smaller step. Therefore, a rejected fragment

conformer in the first step results in more trial conformers in the second step if standard UCONGA is used than if the divide-and-conquer algorithm is used. However, when the standard UCONGA conformer ensemble was investigated, it was found that most of the conformers were generated in the first step (Figure 3.10). This explains the identical ensemble sizes.

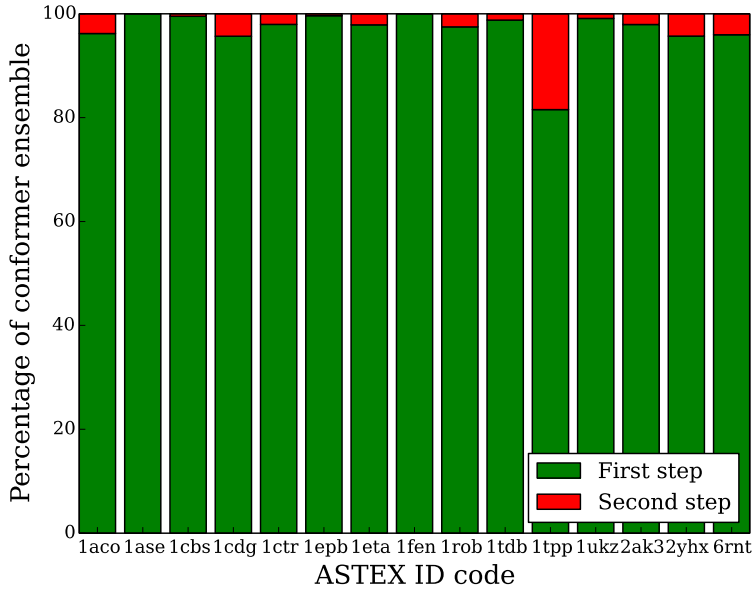



**Figure 3.10** The percentages of the conformer ensemble made up of conformers created in the first and second steps of the standard UCONGA process for the molecules studied in Section 3.4.1.

### 3.4.3 Divide-and-conquer on larger molecules

Although it is reassuring that the divide-and-conquer algorithm reproduces the standard conformer ensemble for less flexible molecules, its purpose is to generate conformer ensembles for larger molecules. For these systems, using RMSD as a metric, it has successfully reproduced the ASTEX reference conformer in all cases where standard UCONGA successfully reproduced the reference. For cases not previously studied using standard UCONGA, all but one reference conformer was successfully reproduced.

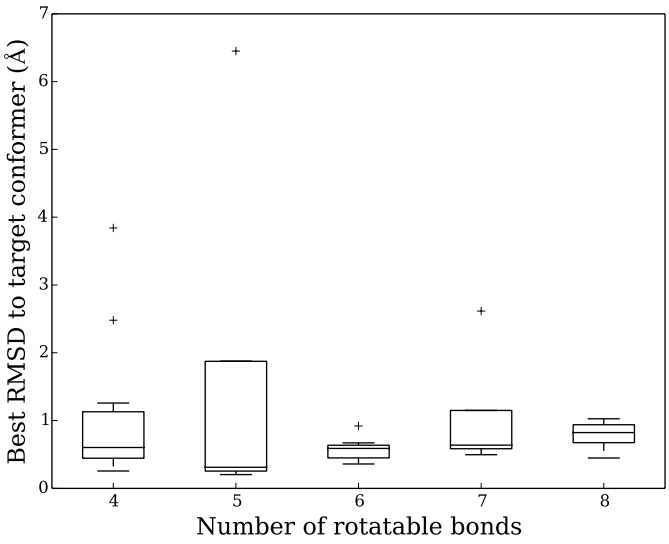



**Figure 3.11** A boxplot showing the RMSD to the reference conformer of the closest conformer in all the ensembles generated using the divide-and-conquer algorithm.

Comparing the RMSD and the torsion-space distance (Figure 3.12), it seems that for these systems RMSD usually finds a better conformer than torsion-space distance, unlike the comparison of the two for standard UCONGA in Section 2.7. This can be explained by comparing the structures of the molecules studied in the two chapters. The molecules studied in Chapter 2 using standard UCONGA, where the two metrics differed, typically consisted of a flexible chain with a bulky rigid or semi-rigid group attached to each end. These bulky groups are heavily weighted by RMSD, causing the alignment of the flexible linker to be neglected. By contrast the molecules studied in Chapter 3 only, using the divide-and-conquer algorithm, typically have a central ring with multiple flexible groups attached. These molecules are adequately described using RMSD.

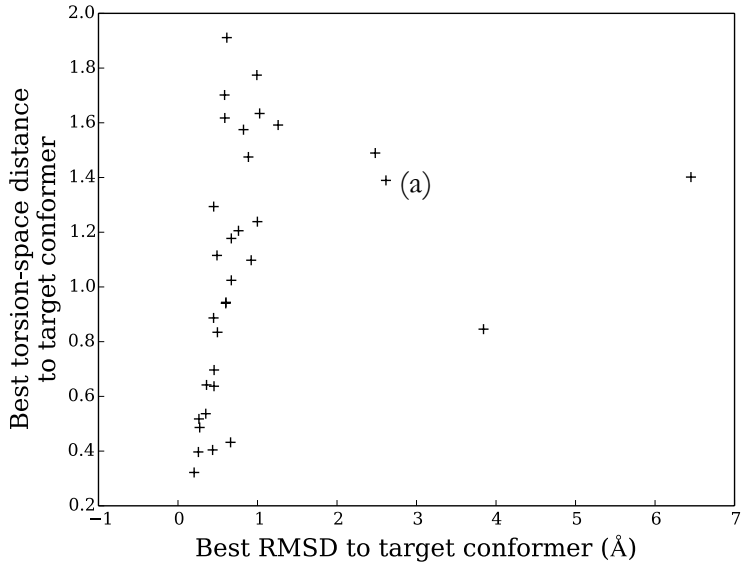



**Figure 3.12** A comparison of the quality of the best conformer by the two metrics, with points of interest labeled.

As was the case with standard UCONGA, when the generated conformer ensembles were filtered using RMSD the result is an approximate ten-fold reduction in ensemble size (Figure 3.13). This suggests that conformer diversity is not reduced by the use of the divide-and-conquer method compared to standard UCONGA.

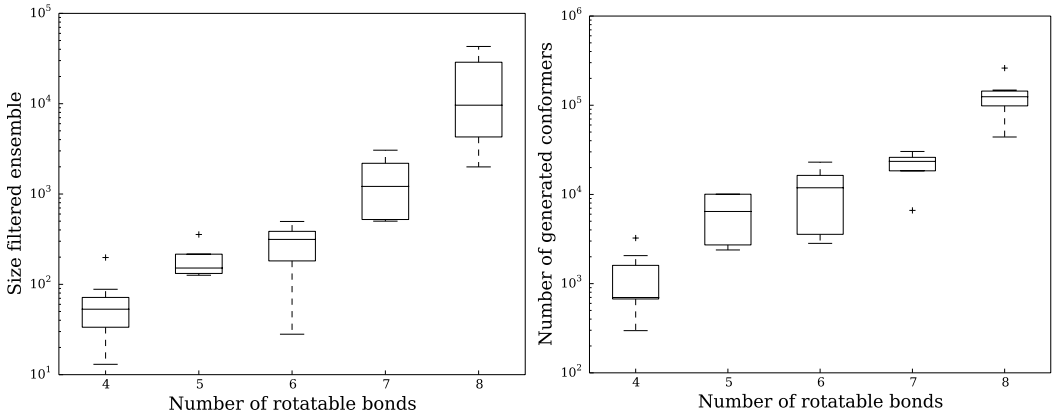



**Figure 3.13** The filtered ensemble size (left) is approximately 10 times lower than the total ensemble size (right).

As mentioned in Section 3.4.2, two of the systems for which the divide-and-conquer algorithm did not generate an acceptable conformer were already studied using the standard UCONGA method, which also failed to generate an acceptable conformer. There is a third system, labelled (a) on Figure 3.12 and C on Figure 3.8, for which no acceptable conformer was generated and which had not already been studied by standard

UCONGA. Investigating this further is useful to determine if there are problems with the divide-and-conquer algorithm, or if the failure to generate a conformer within 2.0 Å of the reference is due to a known weakness of standard UCONGA.

This system is the antibiotic chloramphenicol. The conformers identified as the closest by RMSD and torsion-space distance are similar and visually appear to be a reasonable fit (Figure 3.14), except for the failure to conjugate the nitro group with the benzene ring. This is not due to the divide-and-conquer method, but rather is an inherent problem is a problem inherent with using a method that only considers sterics and ignores electronics. It should be rapidly corrected upon further geometry optimization, if performed.

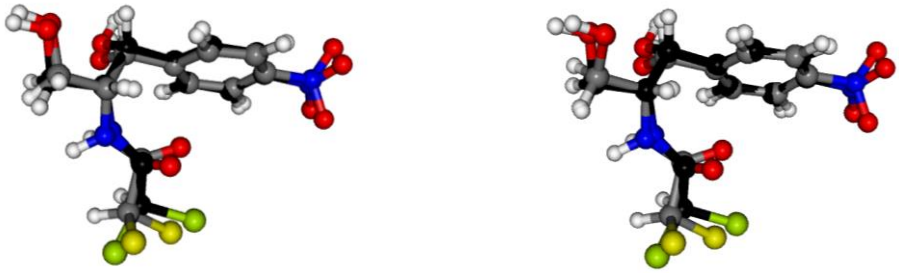


**Figure 3.14** The reference (black carbon skeleton, green chlorine) and best generated (grey carbon skeleton, yellow chlorine) conformers of chloramphenicol according to RMSD (left) and torsion-space distance (right).

### 3.4.4 Computational resources

The purpose of the divide-and-conquer method is to make conformer generation possible for larger molecules, which it has achieved. Estimating the largest molecule for which a conformer ensemble could be generated with UCONGA is difficult. Extrapolating from the work in this chapter is difficult since the size of the conformer ensemble (Figure 3.13) seems to be converging for molecules with up to 7 rotatable bonds, before jumping for molecules with 8 rotatable bonds. This jump is likely due to ring conformers; while molecules with up to 7 rotatable bonds have 0-2 flexible rings, the molecules with 8 have 1-4.

In Section 3.1, it was mentioned that other conformer generation methods could generate conformer ensembles for molecules with up to 12 rotatable bonds. For benchmarking, only molecules with up to 8 rotatable bonds were considered. However, these methods are not directly comparable. One of the methods considered, CONFAB [4], does not generate ring conformers, unlike UCONGA. As discussed in Section 3.3,

ring conformers significantly increase the ensemble size, meaning the performance cannot be compared. Other methods were FROG [5], a Monte Carlo method and BALLOON [6], a genetic algorithm. As discussed in Section 1.1, these methods have lower runtime for large systems at the cost of an increased chance of failing to generate the global minimum, as it is possible to fix the number of conformers they generate and hence the time taken to generate them. In Chapter 7, the divide-and-conquer algorithm is used to generate a conformer ensemble for a molecule with ten rotatable bonds, but it has no flexible rings, possesses nuclear permutational symmetry and has more steric crowding than molecules from the ASTEX data set. Estimating a size limit on the molecules UCONGA can generate conformers for is difficult because, while it can generate conformers for any system, its ability to do so efficiently does vary significantly with the chemical structure.

## 3.5 Conclusion

A divide-and-conquer algorithm has been implemented in the UCONGA package that, once differences in ring conformer handling are taken into account, increases its size limits to be comparable to other deterministic conformer ensemble generation techniques. The use of RMSD filtering of the generated ensemble continues to be useful for reducing the ensemble size at the cost of increasing the runtime. For cases where both standard UCONGA and the divide-and-conquer algorithm were used, the divide-and-conquer algorithm reproduced the standard conformer ensemble. Restricting ring conformer flexibility also helped reduce the runtime and generated ensemble size, at the cost of decreasing the ensemble diversity. This cost is much less for six-membered rings than for five-membered rings.

## 3.6 References


1) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52* (5), 1146–1158.
2) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50* (4), 572–584.
3) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50* (4), 534–546.
4) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.
5) Miteva, M. A; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38* (Web Server issue), W622–W627.
6) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, 47, 2462–2474.

# Chapter 4

## *Ab Initio* benchmarking

## 4.1 Introduction

When performing a computational study using UCONGA, the conformers generated must have their geometries optimized using some other method before their properties of interest are calculated. This is because UCONGA is not a true conformer generation method but rather a sterically-allowed conformation generation method. The structures UCONGA generates are not local minima in the torsional potential energy surface, let alone the global potential energy surface, but are instead structures that are sufficiently sterically unhindered that they should be close to a local minimum. To obtain actual conformers which can be used to interpret experimental data, or for further computations such as reaction mechanism studies or prediction of spectra, these structures must have their geometries optimized. This chapter is concerned with choosing an appropriate method for calculating energies and gradients of the energy during the geometry optimization process, incurring an acceptable trade-off between accuracy and computational cost.

The computationally cheapest methods for evaluating the molecular energy are molecular mechanical forcefields [1] which model molecules using classical mechanics, for instance treating a bond stretching as a spring stretching, but the design of UCONGA is focused toward application to molecules for which there are no good available forcefields. Therefore, they are not a sensible choice for these case studies. The alternative is *ab initio* quantum chemical methods which, by definition, require no *a priori* parameterization. Within this class, there are a range of models of differing speed and accuracy. Hartree-Fock theory (HF) is the cheapest general method, however it does not account for electron correlation, which is responsible for, among other things, London dispersion forces. Density functional theory (DFT) is a family of methods that typically are more accurate than HF at the cost of a higher run-time. Many of them, including the common B3LYP and M06 functionals, are parameterized based on reference data sets so universal accuracy cannot be guaranteed, although they are at least universally applicable. Second-order Moller-Plesset perturbation theory (MP2) is yet more accurate and more expensive, and coupled-cluster theory is highly accurate but too expensive for our purposes, scaling with the sixth power of the number of electrons.

In addition to the methods a basis set, a collection of Gaussian functions used to describe the electron density, must be chosen. The larger this is, the more accurate the

calculations, but the longer it will take. The smallest basis set is STO-3G [2-4], but it lacks the flexibility for atomic orbitals to change shape or size in response to chemical environment. The smallest basis set that allows for this is the common 6-31G* basis set [5-6].

## 4.2 Methods

Five geometry optimization methods were tested: HF/STO-3G [2-4, 7], HF/6-31G* [5-7], B3LYP/6-31G* [5-6, 8], M06/6-31G* [5-6, 9] and MP2/6-31G* [5-6, 10-11].

The methods were tested using the aluminum isopropoxide conformer ensemble that is discussed further in Chapter 7. All 75 of these conformers were optimized using all of these methods, mostly with the QCHEM [12] package but the HF/STO-3G and MP2/6-31G* calculations were performed using GAMESS [13-14]. The default SG-1 integration grid within Q-Chem was not used for the M06 calculations as it can give significant energy errors for this family of functionals [15]. Instead, a Lebedev angular grid with 302 points was combined with a 75-point Euler-MacLaurin radial scheme. The MP2/6-31G* calculations, expected to be the most accurate, were treated as a reference and all other results compared to them. The results of each method were compared with the MP2/6-31G* benchmark based on both geometry and relative energy. Geometries were compared using both RMSD and the maximum torsion angle distance, calculated as

$$\max(\varphi_1^{\text{opt}} - \varphi_1^{\text{unopt}}, \varphi_2^{\text{opt}} - \varphi_2^{\text{unopt}}, \dots, \varphi_n^{\text{opt}} - \varphi_n^{\text{unopt}})$$

The maximum torsion angle distance was used instead of the total torsion-space distance in *k*-means clustering because the torsion-space distance does not distinguish between multiple small bond rotations that largely preserve the reference conformation and one large rotation that produces a different conformer. For instance, total torsion-space distance treats six torsions each twisting by 10° being as great a difference as one torsion twisting by 60° and five others staying constant. When determining if a conformer is the same, these two situations are obviously not equivalent – a small change along many dimensions is unlikely to transition between energy basins on the global potential energy surface while a large change along one dimension might.

## 4.3 Results and discussion

Figure 4.1 illustrates the range of RMSD difference between reference MP2/6-31G* optimized geometries and those obtained by more approximate methods. By minimum, median and maximum RMSD, M06/6-31G* gives the closest geometries to the MP2/6-31G* reference. B3LYP/6-31G* geometries are nearly as good, with quite low RMSDs except for four outliers. Regardless of the basis set, Hartree-Fock geometries have noticeably higher RMSDs to the MP2/6-31G* reference than for either of the density functional methods. Although a RMSD of 1.2 Å was considered acceptable for conformer location with UCONGA, those conformers had not undergone geometry optimization so would not be expected to align as closely as when comparing optimized conformers as is the case here.

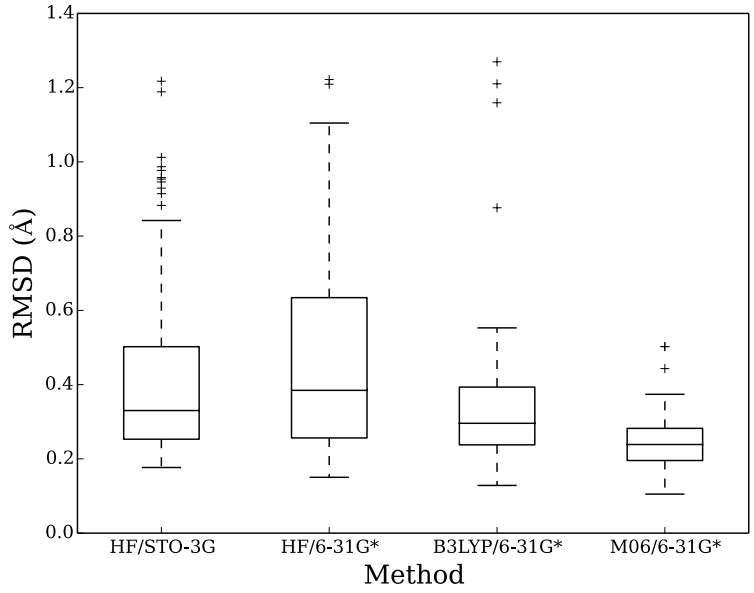



**Figure 4.1** The optimized geometries of 75 aluminum isopropoxide conformers compared to their MP2/6-31G* reference geometries using RMSD.

The same trend is found when the geometries are compared instead using the maximum torsion difference (Figure 4.2). M06/6-31G* has the lowest values, indicating that no conformer has moved too far from the MP2/6-31G* benchmark. B3LYP/6-31G* is nearly as low; that is, nearly as good, except for a handful of outliers. Hartree-Fock theory gives the least similar conformers, with several having at least one torsion angle more than 100° from the MP2/6-31G* benchmark. Regardless of the basis set, HF does not find the same conformers as MP2, making it unsuitable for use in finding low-energy conformers in these case studies.

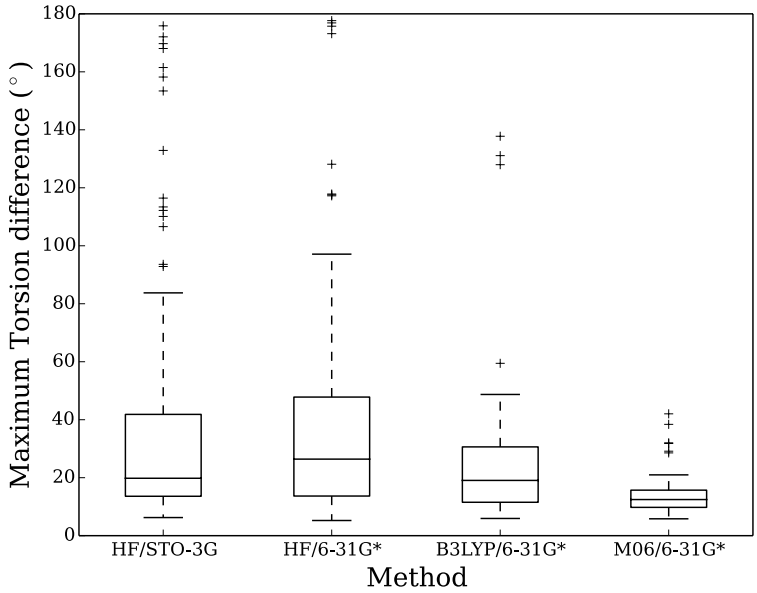



**Figure 4.2** The maximum torsion angle change for 75 aluminum isopropoxide dimer conformers at each level of theory when compared with the MP2/6-31G* reference.

The energies are compared using the absolute difference in relative energies, calculated as

$$\left|\left(E_{MP2,i} - E_{MP2,ref}\right) - \left(E_{L,i} - E_{L,ref}\right)\right|$$

In this equation, $E_{MP2,i}$ is the energy of conformer *i* at the MP2/6-31G* level of theory, $E_{MP2,ref}$ is the energy of the MP2/6-31G* minimum-energy conformer at the MP2/6-31G* level of theory, $E_{L,i}$ is the energy of conformer *i* at level of theory *L*, and $E_{L,ref}$ is the energy of the MP2/6-31G* minimum-energy conformer reoptimized at level of theory *L*. These are illustrated in Figure 4.3. M06/6-31G* performs the worst, with very large relative energy differences. If these were used in chemical applications, the Boltzmann population and hence conformer averages would be very different to the reference MP2/6-31G* values. As for the geometry benchmarking, B3LYP/6-31G* agrees reasonably with the MP2 results, with 75% of conformers within 5 kJ $mol^{-1}$ of their reference values. Once again, Hartree-Fock is not particularly accurate, with at least 25% of conformer energies outside of chemical accuracy, that is, more than 5 kJ $mol^{-1}$ different from the MP2/6-31G* reference. Interestingly, the supposedly superior 6-31G* basis set gives worse energies, with nearly 30% of conformers outside of chemical accuracy, compared with 26% for the STO-3G basis set, presumably due to cancelation of errors.

The poor performance of the M06 functional in calculating relative energies is unexpected and so warrants further discussion. There is some evidence in the literature that the M06 functional can fail to reproduce trends in relative energies, [16] although other studies suggest that it recovers relative energies of stable conformers on the global PES to the same accuracy as MP2 [17]. There are a number of possible explanations including: remaining integration grid discreteness errors, basis set incompleteness errors, implementational issues or parameterization failure. Although a larger than default integration grid was used on each atomic center, this grid may still not be sufficient for describing the electron density associated with the aluminum atoms, as grids are typically developed primarily for organic molecules. Basis set incompleteness is also a realistic problem, as M06 was parameterized using triple-zeta basis sets [9], while only a double-zeta basis set was used for these calculations. The complexity of the M06 functional [9] increases the chances of implementation error, but also the propensity to fail outside the chemical space over which the functional was parameterized. The $Al_2O_2$ ring is an uncommon structural motif that is unlikely to lie within the parameterization space of the M06 functional [9].

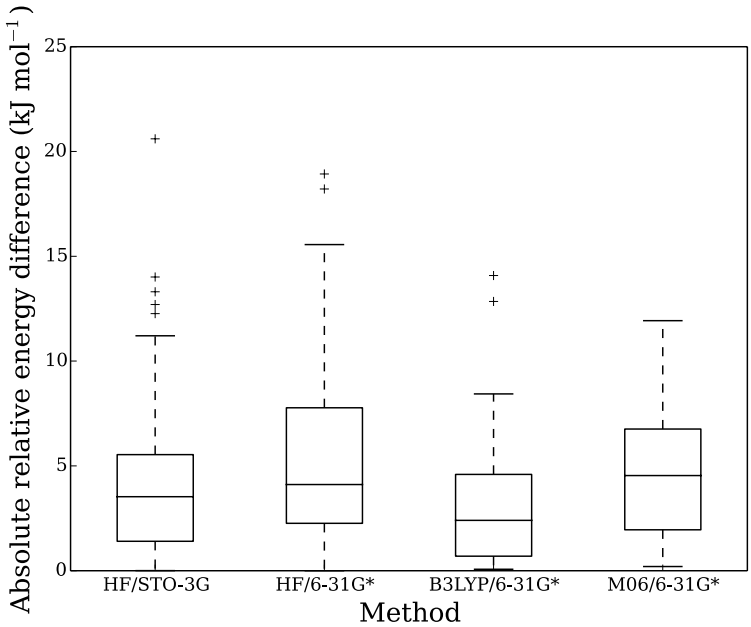



**Figure 4.3** The relative energies of the conformer ensemble compared to the reference calculated using MP2/6-31G*

Finally, considering the run times summarized in Figure 4.4, it is clear that performing full MP2/6-31G* geometry optimization for all conformers in the case studies is likely to consume an impractically large amount of time. Some geometry optimizations took over

11 days of real time, compared to just over 1 day for the longest M06/6-31G* optimization. In general, runtime decreases with method sophistication as expected, although it is interesting that two HF/6-31G* optimizations took more time than any of the M06/6-31G* or B3LYP/6-31G* calculations. These calculations required many steps due to these conformers not being stable according to the HF/6-31G* method, further evidence of its unsuitability for these calculations.

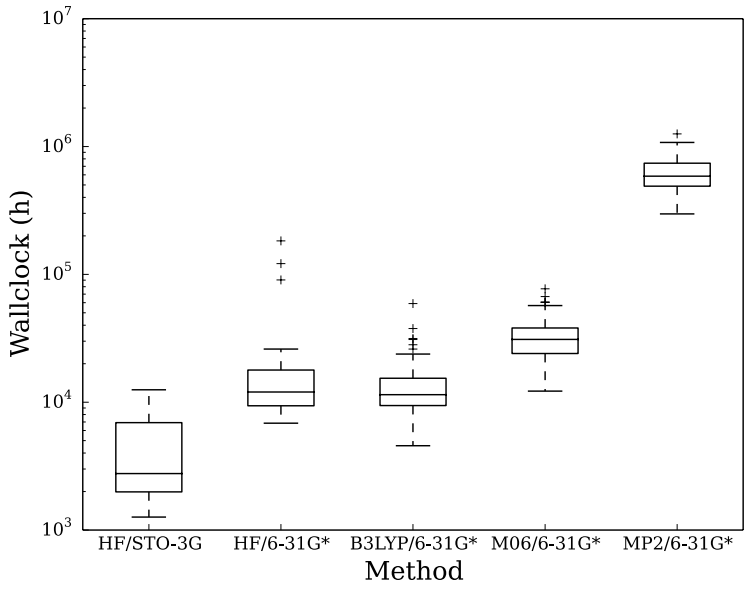



**Figure 4.4** The distribution of run-times needed for geometry optimization with various methods, including the MP2/6-31G* reference.

### 4.3.1 Investigation of B3LYP outliers

From the distributions alone, it can be concluded that M06 can successfully reproduce the reference geometries but not their energies, and HF can reproduce neither regardless of the basis set. B3LYP/6-31G* is a more complicated case. On average, it reproduces the reference geometries nearly as well as M06/6-31G* and, on average, it is the only method to successfully reproduce the energies, but there are five outlying conformers (one of the four outliers by RMSD is the same as one of the two outliers by energy) where B3LYP/6-31G* fails to reproduce the benchmark geometry, energy or both. As was the case for benchmarking UCONGA (Section 2.7), investigating outliers may help to understand the limitations of the method, namely B3LYP/6-31G*. These outliers are presented in Table 4.1. The reference energy is presented, along with the RMSD, maximum torsion difference and energy difference mentioned previously.

**Table 4.1** Important statistics for conformers that changed significantly relative to the MP2/6-31G* reference on optimization at the B3LYP/6-31G* level of theory.

| ID | ΔE (kJ mol$^{-1}$)[a] | E (MP2) (kJ mol$^{-1}$)[b] | RMSD (Å) | Maximum torsion difference (°) |
|---|---|---|---|---|
| 1 | -14.5 | 17.3 | 1.3 | 137 |
| 2 | -13.3 | 18.9 | 0.2 | 131 |
| 3 | 0.17 | 6.2 | 1.2 | 36 |
| 4 | 3.8 | 4.5 | 1.2 | 20 |
| 5 | 4.1 | 4.5 | 0.9 | 10 |

[a] ΔE = E(MP2) – E(B3LYP)
[b] E(MP2) is the MP2/6-31G* energy relative to the global minimum.

These systems fall into two groups: rows 1 and 2 (high energy conformers which undergo significant torsion rotation to lower their energy) and rows 3-5 (no single torsion angle changes significantly upon optimization.

Outliers 1 and 2 are different conformers in the reference MP2/6-31G* and B3LYP/6-31G* ensembles. However, the reference conformers will not be significantly populated at room temperature or even 500 K due to their high energies. The fact that they collapse upon optimization with B3LYP/6-31G* to a more stable conformer is therefore not a particular problem. The fact that they collapse readily when a lower level of theory is used suggests that they are shallow minima, possibly stabilized by dispersion which B3LYP does not capture at all.

Outliers 3-5, by contrast, are still the same conformer in the reference and B3LYP/6-31G* ensembles. The energies are relatively low and changed relatively little. While some torsions have changed by 10°-30°, they are still in the same local minimum for rotation around that particular bond. The bounding boxes of these conformers have not changed significantly either. The large change in the RMSD is because these are large molecules – 8 bonds, 4 of which are considered rotatable by UCONGA, so small changes in the torsions can create relatively large changes in the atomic coordinates. However, the analysis methods used in the case studies rely primarily on torsion angles and bounding boxes. Atomic coordinates are only considered for pre-optimization filtering. These conformers are therefore considered sufficiently similar to their reference geometries, although this would not be the case were they being used for an application dependent on atomic coordinates such as gas-phase electron diffraction structure refinement.

The fact that conformer 2 from the set of B3LYP/6-31G* outliers has a large torsion angle difference but low RMSD difference requires further investigation, because

typically both RMSD and torsion angle differences are both used as indicators of conformer quality and/or agreement. If RMSD is not a reliable metric of conformer similarity, it should not be used to filter large conformer ensembles or as a metric of diversity. Investigation reveals that this low RMSD value arises from the majority of the molecule being very well aligned with a single misaligned isopropyl group. This failure to distinguish between different conformers by RMSD is only likely to occur for large molecules with a central core and many flexible peripheral groups. This averaging out of one small very-poorly-fitting section in a large molecule that is otherwise well-aligned is a weakness that must be accepted for any size-independent metric of conformer similarity. However, throughout the rest of this thesis, a range of metrics and analysis methods will be deployed simultaneously, as they provide complementary information and together provide a robust picture of differences and similarities between conformers.

## 4.4 Conclusion

When runtime is traded against accuracy, optimizing the geometries at the M06/6-31G* level of theory, then subsequently calculating the energy using MP2/6-31G*, is the best available option. For simple, less complicated systems, performing a simple B3LYP/6-31G* optimization is acceptable.

More generally, this work serves as a reminder of the importance of testing computational methods, especially density functionals, before use. Highly parameterized density functions, while highly accurate most of the time, can break down unpredictably, especially when calculating relative energies. Geometry optimizations are less prone to this failure of parameterization than energy calculations are, but confirmation of accuracy is nevertheless important.

## 4.5 References


1) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons, **2013.**
2) Hehre, W. J.; Stewart, R. F.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **1969**, *51* (6), 2657–2664.
3) Hehre, W. J.; Ditchfield, R.; Stewart, R. F.; Pople, J. A. Self-Consistent Molecular Orbital Methods. IV. Use of Gaussian Expansions of Slater-Type Orbitals. Extension to Second-Row Molecules. *J. Chem. Phys.* **1970**, *52* (5), 2769–2773.
4) Gordon, M. S.; Bjorke, M. D.; Marsh, F. J.; Korth, M. S. Second-Row Molecular Orbital Calculations. 5. A Minimal Basis INDO for Na-Cl. *J. Am. Chem. Soc.* **1978**, *100* (9), 2670–2678.
5) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta.* **1973**, *28* (3), 213–222.
6) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665.
7) Roothaan, C. C. J. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23* (2), 69–89.
8) Stephens, P. J.; Devline, F. J.; Chablowski, C. F.; Frisch, M. J. *Ab Initio* Calculation of Vibrational Absorption and Circular Dichrroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627.
9) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2007**, 120 (1-3), 215–241.
10) Möller, C.; Plesset, M. S. Note on an Approximation for Many-Electron Systems. *Phys Rev*, **1934**, *240* (1), 618–622.
11) Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. A Derivation of the Frozen-Orbital Unrestricted Open-Shell and Restricted Closed-Shell Second-Order Perturbation Theory Analytic Gradient Expressions. *Theor. Chim. Acta.* **2003**, *110* (4), 233–253.
12) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L.

D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, A.; Steele, R. P.; Sundstrom, E. J.; Woodcock III, H. L.; Zimmerman, M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, B.; Brown, S. T.; Casanova, D.; Chang, C.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Jr, R. A. D.; Do, H.; Dutoi, A. D.; Edgar, G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Philipp, H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V; Levchenko, V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.; Marenich, A. V; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; Neill, P. O.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer III, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Steven, R.; Hsu, C.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Lyudmila, V.; Subotnik, J. E.; Voorhis, T. Van; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M.; Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *111* (2), 184–215.

13) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: GAMESS a Decade Later. In Theory and Applications of Computational Chemistry; **2005**; pp 1167–1189.

14) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14* (11), 1347–1363.

15) Wheeler, S. E.; Houk, K. N. Integration Grid Errors for Meta-GGA-Predicted Reaction Energies: Origin of Grid Errors for the M06 Suite of Functionals. *J. Chem. Theory Comput.* **2010**, *6* (2), 395–404.

16) Chan, B.; Gilbert, A. T. B.; Gill, P. M. W.; Radom, L. Performance of Density Functional Theory Procedures for the Calculation of Proton-Exchange Barriers: Unusual Behavior of M06-type Functionals. *J. Chem. Theory Comput.* **2014**, *10* (9), 3777–3783.

17) Fu, Z.; Li, X.; Merz Jr, K. M. Conformational Analysis of Free and Bound Retinoic Acid. *J. Chem.Theory Comput.*, **2013**, *125* (21), 2621–2629.

# Chapter 5

## Case study: Sterically crowded molecules

## 5.1 Introduction

Sterically crowded molecules pose interesting conformational challenges. Despite their many bonds, their conformer ensembles are often small. These conformers often adopt unusual torsion angles to allow for the steric crowding, causing problems for knowledge-based conformer generation methods. For instance, as discussed in Chapter 2, 1,1,2-tri-*tert*-butyl disilane [1] favors an eclipsed conformer, minimizing gauche interactions between the bulky *tert*-butyl groups, in preference to the staggered conformer that would be expected to be the minimum.

Sterically crowded molecules are chemically interesting because they feature a delicate balance between repulsive and bonding interactions that can lead to spontaneous homolytic dissociation in the gas phase [2, 3] or a change of oligomerization state upon melting [4]. Steric crowding can lead to a variety of unusual bonding arrangements, particularly torsional alignments, in order to minimize steric repulsion. This makes modeling/determining their structures difficult, especially when second-row elements such as silicon and phosphorus are bonded to first-row elements such as carbon [5].

Structures that are difficult to compute serve as useful benchmarks for computational methods [6]. As, by their very nature, they are hard to compute, a mechanism to obtain good quality experimental structures is needed. This enables structure comparison between results of known computational methods with experiments, and also allows newly-developed methods to be compared to known methods. Gas-phase electron diffraction (GED) [5] is the most widely used technique for experimentally determining gas-phase structures, free from crystal packing effects. More recently, advances in experimental techniques and analysis programs have enabled the study of bulky molecules by GED [5]. However, due to the size of these molecules, *ab initio* calculations are often needed to assist with interpretation of the experimental data [5]. As discussed in Chapter 2, potential energy surface scanning for these large systems is often not possible, therefore UCONGA is expected to generate useful conformer ensembles for this purpose.

A selection of these bulky molecules have been chosen as test cases for UCONGA (Figure 5.1). The first, aluminum isopropoxide tetramer (AIPT), is extremely large with 18 rotatable bonds. However, it has approximately $D_2$ point-group symmetry [7] in

crystalline form and NMR data [4] suggest the same is true in solution, reducing the conformer search space to 3 unique rotatable bonds. As mentioned earlier, 1,1,2-tri-*tert*-butyl disilane (TTBS) adopts an eclipsed conformation at the global minimum, providing a good test of how universal UCONGA is. Tetrakis(trimethylsilyl)diphosphane (TTSP) and hexakis(trimethylsilyl)disilane (HTSS) are related symmetrical bulky systems. Gas-phase electron diffraction data has been collected for them, but their structures have not been determined, so additional *ab initio* geometric and energetic data may help support the structural refinement process.

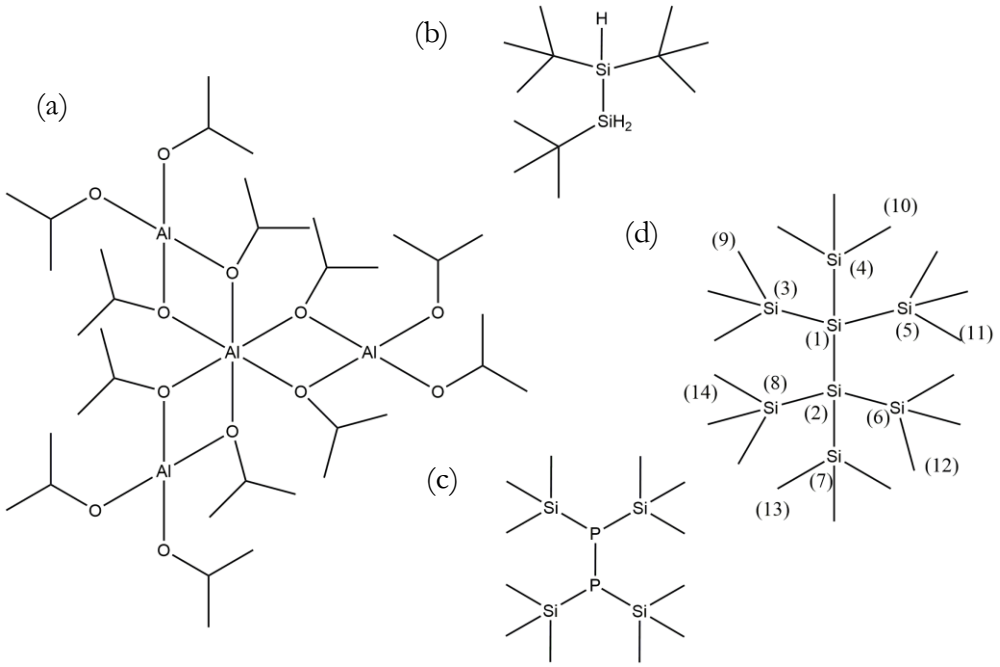



**Figure 5.1** The chemical structures of (a) aluminum isopropoxide tetramer (AIPT), (b) 1,1,2-tri-*tert*-butyl disilane (TTBS), (c) tetrakis(trimethylsilyl)diphosphane (TTSP) and (d) hexakis(trimethylsilyl)disilane (HTSS). For HTSS, where the conformer ensemble is large enough that analysis beyond viewing all of the conformer structures is necessary, torsion labels around rotatable bonds are given in Table 5.1.

**Table 5.1** Bond labels for HTSS

| Torsion | Label |
|---|---|
| C | 3-1-2-6 |
| T11 | 9-3-1-2 |
| T12 | 10-4-1-2 |
| T13 | 11-5-1-2 |
| T21 | 12-6-2-1 |
| T22 | 13-7-2-1 |
| T23 | 14-8-2-1 |

## 5.2 Methods

For TTBS, TTSP and HTSS, conformer ensembles were generated using UCONGA with an initial step size of 60° and a secondary step size of 30°. Due to its size, this is impractical for AIPT. UCONGA was modified to enforce point-group symmetry for this special case; symmetry-equivalent bonds were assigned the same torsion angle. (Figure 5.2). In addition, only the 60° step size was used.

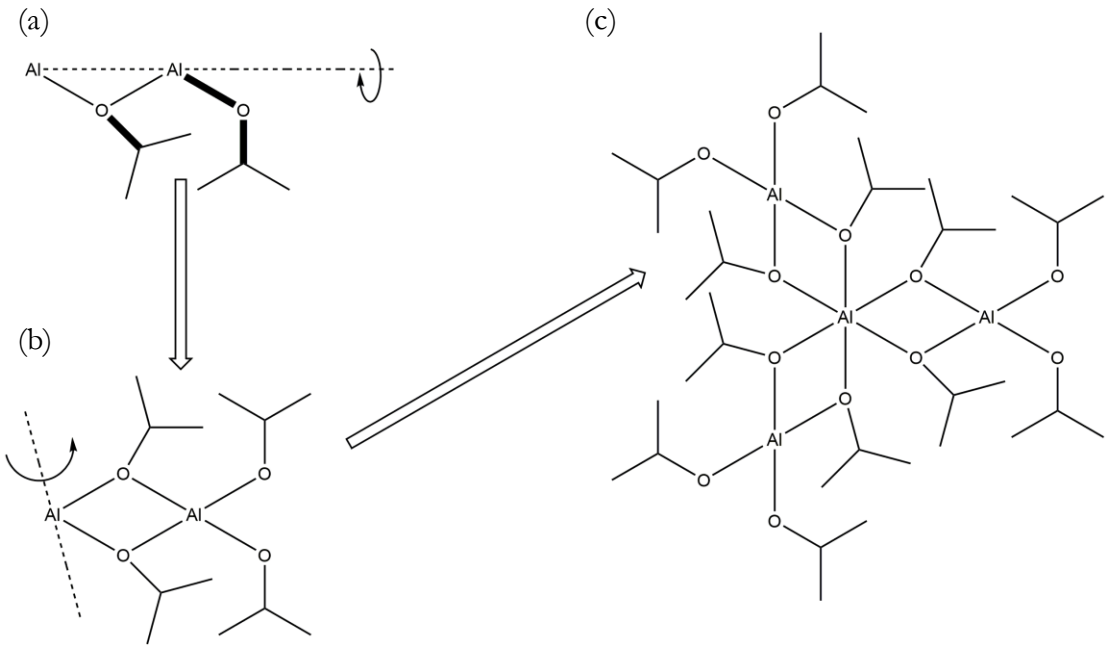



**Figure 5.2** (a) There are only three crystallographically unique rotatable bonds in AIPT (in bold). The modified version of UCONGA only sets the torsion angles around these three bonds. They are related by a $C_2$ axis (dashed line) to the other half of the $Al_2O_2$ ring (b) One $Al_2O_2$ ring is related to the other two by a $C_3$ axis (dashed line) running through the central aluminum atom and the centers of two opposite faces of the octahedron defined by the bridging oxygen atoms. (c) The whole AIPT structure.

The generated conformers were optimized at the M06/6-31G* level of theory [6, 8-9] as implemented in NWChem [10] (AIPT) and Gaussian [11] (TTSP and HTSS), with single point energies at the optimized geometries calculated at MP2/6-31G* [12] using QChem [13]. The generated conformers for TTBS were not optimized since this system had already been studied using GED and so the generated conformers were only needed for comparison with known conformers, not for any further use. The structures of all generated conformers are given in the electronic appendix.

## 5.3 Results and discussion

### 5.3.1 Tetrameric aluminum isopropoxide

Due to the enforced symmetry and extreme steric crowding of this system, UCONGA produced only two sterically-allowed conformers (Figure 5.3). Following geometry optimization, the lowest-energy is similar to the crystal structure, whereas the other is much higher in energy and does not represent a stable minimum energy conformer on the potential energy surface.

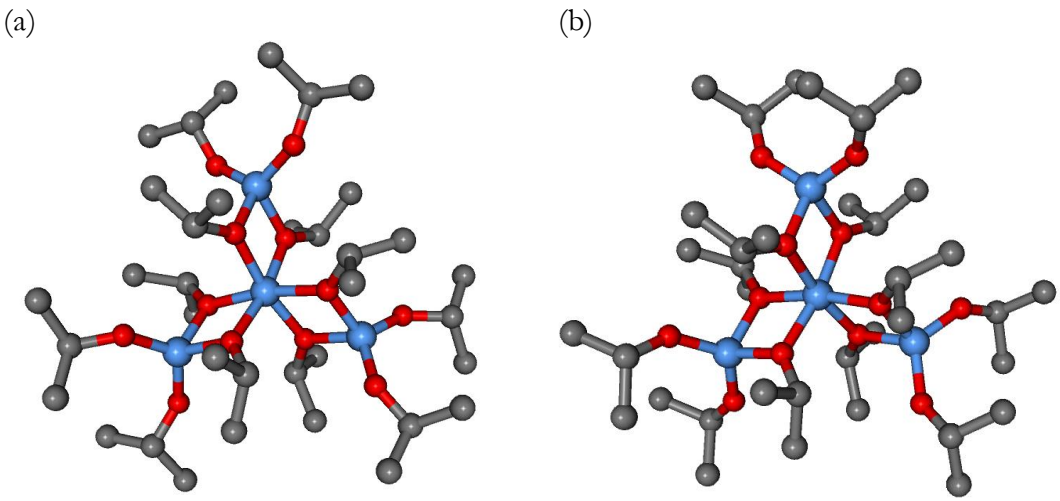
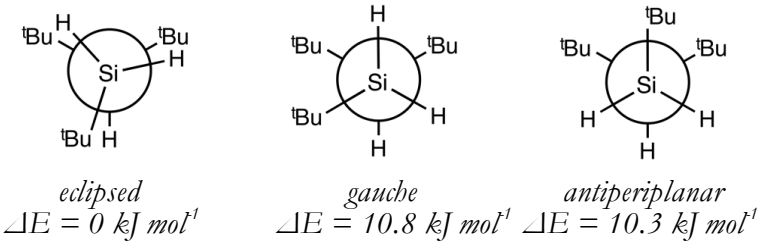



**Figure 5.3** (a) The stable conformer of AIPT. (b) The unstable conformer of AIPT found by UCONGA. Hydrogen atoms have been removed for clarity.

### 5.3.2 1,1,2-tri-*tert*-butyldisilane

As mentioned earlier, the lowest energy conformer found by a previous computational study [1] shows the unique *tert*-butyl group *eclipsed* with the unique proton. The study also found two higher energy conformers, in which the unique *tert*-butyl group and unique proton are *gauche* and *antiperiplanar* (Figure 5.4). These were calculated to be 10.8 kJ $mol^{-1}$ and 10.3 kJ $mol^{-1}$ higher in energy than the *eclipsed* conformer respectively and only contribute 2.3% and 1.4% to the conformer ensemble at room temperature. At the experimental temperature of 411K, they contribute 7.5% and 4.3% each.




*eclipsed* $\Delta E = 0$ kJ mol$^{-1}$ *gauche* $\Delta E = 10.8$ kJ mol$^{-1}$ *antiperiplanar* $\Delta E = 10.3$ kJ mol$^{-1}$

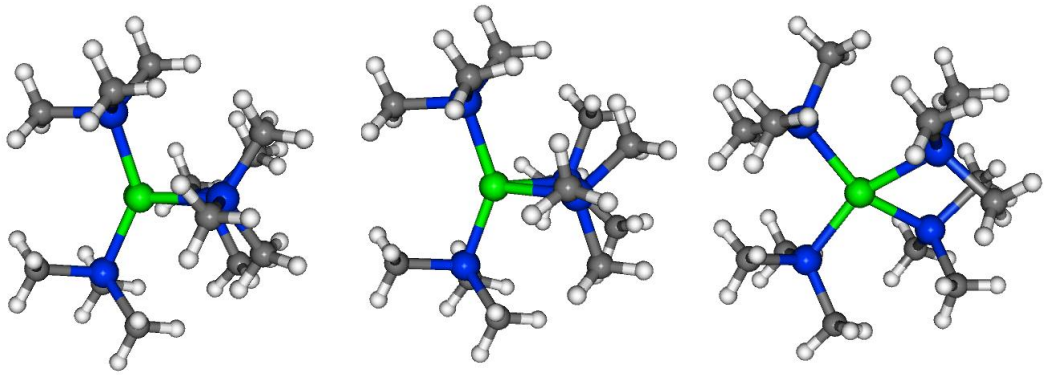
**Figure 5.4** Newman projection representations of 1,1,2-tri-*tert*-butyldisilane conformers viewed down the silicon-silicon bond.

UCONGA found four conformers, including the *eclipsed* and *gauche* conformers. The two other conformers identified by UCONGA resemble the *eclipsed* conformer, only differing in extent of rotation about the silicon – *tert*-butyl bonds, leading to different *tert*-butyl group packing. A finer-grained potential energy surface search was performed with UCONGA, where a step size of 15° was used as well as the 60° and 30° steps. This generated a much larger ensemble, with 1155 conformers, but it still did not include the *antiperiplanar* conformer.

There are two reasons why the *antiperiplanar* conformer is not located. First, the previous experimental and computational data indicate that bond angles are coupled to the torsion angles [1]; the C-Si-Si bond angle involving the unique *tert*-butyl group increases from 113° in the *eclipsed* conformer to 124° in the *antiperiplanar* conformer to reduce the interaction between the two *tert*-butyl groups. Locating this conformer would require a relaxed potential energy surface search; that is, one in which each trial conformer undergoes geometry optimization over all coordinates except the torsions varied by UCONGA, using either a forcefield (with the attendant problems of limited parameterization discussed in Chapters 1 and 2) or an *ab initio* method (with the attendant increase in computational cost). Second, the minimum around the *antiperiplanar* conformer on the potential energy surface is very narrow. If all torsions around rotatable bonds in the previously identified *antiperiplanar* conformer are rounded to the nearest 15°, then even with the increased bond angle two hydrogen atoms come closer than 0.7 times the sum of their van der Waals radii and so would be rejected by UCONGA. All conformer location techniques without relaxation over spectator coordinates can fail to locate local minima if they are too narrow and fall between two search points.

### 5.3.3 Tetrakis(trimethylsilyl)diphosphane

UCONGA initially produced eight sterically-allowed conformations, which converged to three distinct conformers upon optimization (Figure 5.5). Two of the optimized conformers are within 4 kJ $mol^{-1}$ of each other, and the third is 16 kJ $mol^{-1}$ higher in energy than the minimum. The second lowest-energy conformer has previously been identified in the literature at the HF/6-31G* level of theory [14-15] and recent re-investigation of this system additionally identified the lowest-energy conformer [16], based upon constrained potential energy surface scans at the HF/3-21G* and MP2/6-31G* levels of theory. The highest-energy conformer was only found using UCONGA. All were confirmed as local minima by vibrational analysis at the MP2/6-31G* level of theory.



$\Delta E = 0$ kJ $mol^{-1}$
$\varphi_{Si\text{-}P\text{-}P\text{-}Si} = -4°$

$\Delta E = 3.6$ kJ $mol^{-1}$
$\varphi_{Si\text{-}P\text{-}P\text{-}Si} = -11°$

$\Delta E = 15.8$ kJ $mol^{-1}$
$\varphi_{Si\text{-}P\text{-}P\text{-}Si} = -55°$

**Figure 5.5** The three optimized conformers of TTSP aligned in Newman projection format, viewed down the P-P bond.

### 5.3.4 Hexakis(trimethylsilyl)disilane

UCONGA generated 42 conformers that stayed unique upon geometry optimization, more than for TTSP despite the greater steric crowding and symmetry of HTSS. All are within 5 kJ $mol^{-1}$ of the global minimum. Clustering based on the torsion angles (Figure 5.6) separates the conformer ensemble into two indistinct, similarly-sized clusters. The cluster centers are identical for most torsion angles. For those torsion angles where the cluster centers are different the clusters still overlap. Both clusters contain low-energy conformers, with Boltzmann-averaged cluster energies of 0.22 and 0.17 kJ $mol^{-1}$. Clustering using the bounding box approach (Figure 5.6) also identifies two similarly-sized clusters (most central conformers shown in Figure 5.7). The bounding-box clusters

are distinguished according to the length of the molecules along the tertiary (longest) axis of inertia, although they are adjacent to each other with no gap. The tertiary axis of inertia has the smallest associated moment of inertia and therefore is usually the longest span of the molecule. They are not separated on the secondary and primary axes of inertia. These are visually quite similar to each other, which is consistent with the similar energies and overlapping cluster centers. However, there are two notable outliers along the primary axis of inertia, corresponding to more compact structures relative to the others, which are illustrated in Figure 5.8.

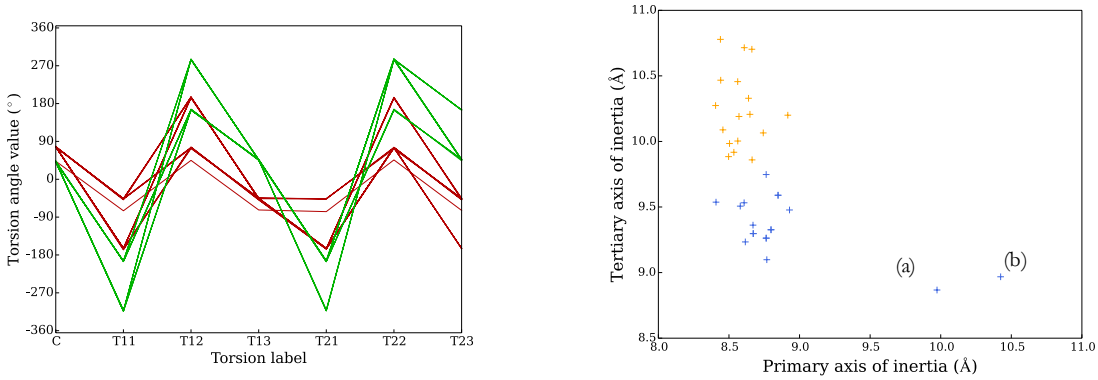



**Figure 5.6 Left** A parallel coordinates plot of the torsion angles of the HTSS conformer ensemble, colored according to torsion-based cluster. Bond labels are given in Table 5.1. **Right** A scatter plot of the axes of inertia of the HTSS conformer ensemble, colored according to bounding-box cluster.

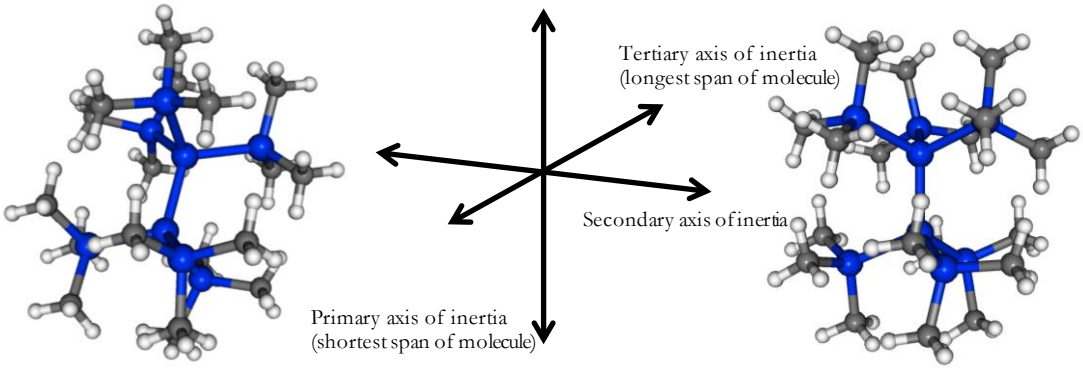



**Figure 5.7** The molecular structures, viewed down the central bond, of the most central conformers of the blue (left) and orange (right) bounding-box clusters.

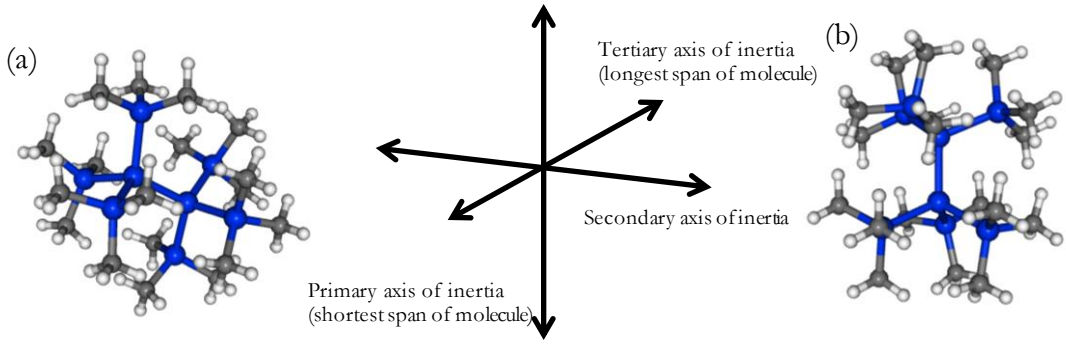



**Figure 5.8** The structures of the two conformers with unusually long primary axes of inertia, labeled (a) and (b) on the right panel of Figure 5.6 aligned as for Figure 5.7.

The bounding-box clustering did locate one conformer where the tertiary axis of inertia significantly deviates from the central Si-Si bond. This may be important for any further study where molecular rotation is important. However, for both forms of clustering, the differences between clusters mostly seem to come from the additive effects of many small changes instead of a few large ones, making the usefulness of cluster analysis for this system limited.

While the conformer ensemble is quite large, many of the conformers are quite similar to each other by both metrics, so gas-phase electron diffraction studies may be possible, using the lowest-energy or most central conformers from each torsional or bounding-box cluster. If this provides a poor representation of the data, one or both of the bounding-box outliers shown in Figure 5.8 could be included as well, or the conformers most dissimilar to those already included.

## 5.4 Conclusion

This set of molecules illustrates the power of using nuclear permutational symmetry to reduce the number of conformers generated. For instance, TTSP has five rotatable bonds. The smallest conformer ensemble for a molecule of that size in the less symmetrical ASTEX test set is 1369 conformers, compared with only eight for TTSP. While some of that is due to the greater steric bulk of TTSP leading to more trial conformers being rejected, the contribution of symmetry is also important. Since each of the four *tert*-butyl groups has threefold rotational symmetry, the ensemble size would otherwise be $3^4 = 81$ times bigger if permutational symmetry was not used. UCONGA has generated conformers that existing methods such as reduced-dimensional relaxed

potential energy surface searches cannot discover due to their inability to simultaneously search all rotatable bonds. However, it cannot find minima for systems with a high degree of coupling between torsion and bond angles.

## 5.5 References


1) Hinchley, S. L.; Smart, B. A.; Morrison, C. A.; Robertson, H. E.; Rankin, D. W. H.; Zink, R.; Hassler, K. 1,1,2-Tri-*Tert*-Butyldisilane, $Bu^t_2HSiSiH_2Bu^t$: Vibrational Spectra and Molecular Structure in the Gas Phase by Electron Diffraction and Ab Initio Calculations. *J. Chem. Soc., Dalton Trans.* **1999**, 2303–2310.
2) Hinchley, S. L.; Morrison, C. A.; Rankin, D. W. H.; Macdonald, C. L. B.; Wiacek, R. J.; Cowley, A. H.; Lappert, M. F.; Gundersen, G.; Clyburne, J. A. C.; Power, P. P. Persistent Phosphinyl Radicals from a Bulky Diphosphine: an Example of a Molecular Jack-in-the-Box. *Chem. Commun.*, **2000** (20), 2045–2046.
3) Hinchley, S. L.; Morrison, C. A.; Rankin, D. W. H.; Macdonald, C. L. B.; Wiacek, R. J.; Voigt, A.; Cowley, A. H.; Lappert, M. F.; Gundersen, G.; Clyburne, J. A. C.; Power, P. P. Spontaneous Generation of Stable Pnictinyl Radicals from "Jack-in-the-Box" Dipnictines: a Solid-State, Gas-Phase, and Theoretical Investigation of the Origins of Steric Stabilization. *J. Am. Chem. Soc.* **2001**, *123* (37), 9045–53.
4) Shiner Jr, V. J.; Whittaker, D.; Fernandez, V. P. The Structures of Some Aluminum Alkoxides. *J. Am. Chem. Soc.* **1963**, *2* (6), 2318–2322.
5) Mitzel, N. W.; Rankin, D. W. H. SARACEN – Molecular Structures from Theory and Experiment: The Best of Both Worlds. *Dalton Trans.* **2003**, (19), 3650–3662.
6) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2007**, *120* (1-3), 215–241.
7) Turova, N. Y.; Kozunov, V. A.; Yanovskii, A. I.; Bokii, N. G.; Struchkov, Y. T.; Tarnopol'skii, B. L. Physico-Chemical and Structural Investigation of Aluminium Isopropoxide. *J. Inorg. Nucl. Chem.* **1979**, *41* (1), 5–11.
8) Francl, M. M.; Pietro, Wi. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665.
9) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222.
10) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Dam, H. J. J. Van; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; DeJong, W. A. NWChem : A

Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations. *Comput. Phys. Commun.* **2010**, *181* (9), 1477–1489.

11) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V; Cioslowski, J.; Fox, D. J. Gaussian09 Revision D.01, **2009**.

12) Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. A Derivation of the Frozen-Orbital Unrestricted Open-Shell and Restricted Closed-Shell Second-Order Perturbation Theory Analytic Gradient Expressions. *Theor. Chim. Acta* **2003**, *110* (4), 233–253.

13) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, A.; Steele, R. P.; Sundstrom, E. J.; Woodcock III, H. L.; Zimmerman, M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, B.; Brown, S. T.; Casanova, D.; Chang, C.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Jr, R. A. D.; Do, H.; Dutoi, A. D.; Edgar, G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Philipp, H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V; Levchenko, V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.; Marenich, A. V; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; Neill, P. O.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Rehn, D.

R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer III, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Steven, R.; Hsu, C.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Lyudmila, V.; Subotnik, J. E.; Voorhis, T. Van; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M.; Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *111* (2), 184–215.

14) Tekautz, G.; Hassler, K. Molecular Inorganic Silicon Chemistry: Conformational Properties of 1,1,2,2-Tetrakis(trimethylsilyl)disilanes and of Tetrakis(trimethylsilyl) Diphosphine: A Comparative Vibrational Spectroscopic and Ab Initio Study. In Organosilicon Chemistry VI: From Molecules to Materials; **2008**; pp 368–372.

15) Borisenko, K. B.; Rankin, D. W. H. Structural Changes, P − P Bond Energies, and Homolytic Dissociation Enthalpies of Substituted Diphosphines from Quantum Mechanical Calculations. *Inorg. Chem.* **2003**, *42* (22), 7129–7136.

16) Humphrey-Taylor, H. Determination of the Gas Phase Structures of Tetrakis(trimethylsilyl)diphosphine and Hexakis(trimethylsilyl)disilane. Honours Thesis, University of Canterbury, **2014**.

# Chapter 6

## Case study: Molecules on surfaces

## 6.1 Introduction

While UCONGA is designed to be a universal method, most of the case studies described so far have used UCONGA to study the gas-phase structure of sterically crowded branched molecules with a p-block core, while most of the benchmarking was performed on protein-bound ligands. There are many interesting molecules not included in either of those categories.

One such group of molecules, which is different to those used in the previous case-studies and benchmarking, and whose activity is affected by its conformational properties, are some redox-active surface-modification agents recently studied by the Downard group at the University of Canterbury (Figure 6.1) [1]. These include *p*-nitrobenzylamine (NBA), 1-(ferrocenylmethyl) 4-phenyl-1,2,3-triazole (FMPT), *N*-benzyl *p*-nitrobenzamide (BNB), *N*-benzyl ferrocenylacetamide (BFA) and *N*-(ferrocenylmethyl) benzamide (FMBA).

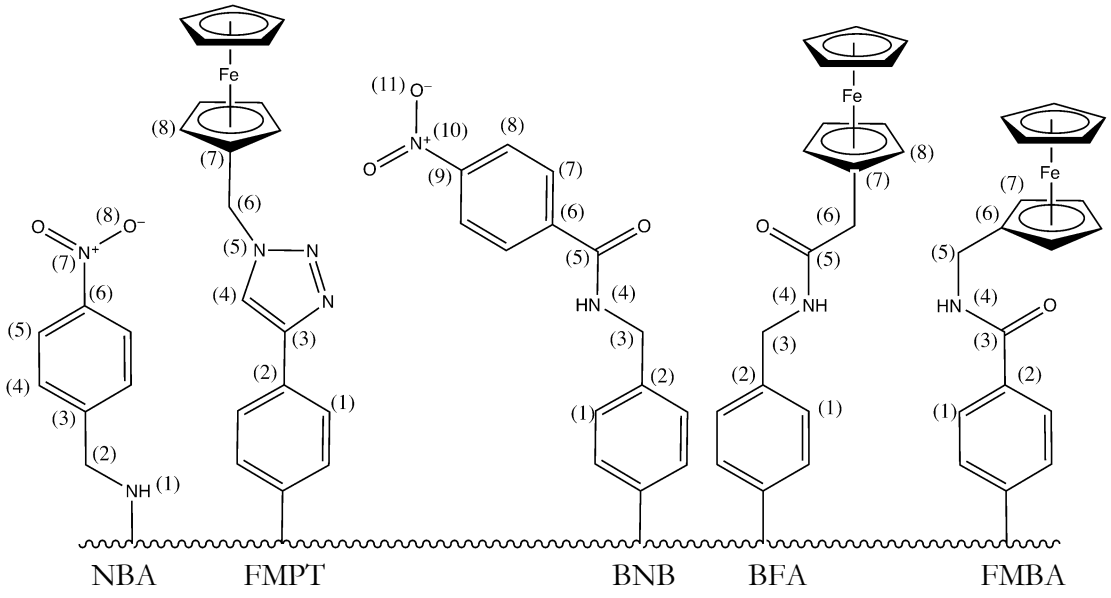



**Figure 6.1** Five surface modification agents bound to an arbitrary surface. Torsion atoms about rotatable bonds are defined by atom connectivities listed in Table 6.1.

**Table 6.1** Torsion angle labels for rotatable bonds of molecules in Figure 6.1.

| Molecule | Torsion Angle Labels | | | | |
|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** |
| NBA | 1-2-3-4 | 5-6-7-8 | N/A | N/A | N/A |
| FMPT | 1-2-3-4 | 4-5-6-7 | 5-6-7-8 | N/A | N/A |
| BNB | 1-2-3-4 | 2-3-4-5 | 3-4-5-6 | 4-5-6-7 | 8-9-10-11 |
| BFA | 1-2-3-4 | 2-3-4-5 | 3-4-5-6 | 4-5-6-7 | 5-6-7-8 |
| FMBA | 1-2-3-4 | 2-3-4-5 | 3-4-5-6 | 4-5-6-7 | N/A |

This group of molecules forms an interesting case study for three reasons. First, the ferrocenyl motif in BNB, BFA and FMBA is common but not included in the benchmarking dataset or the other case studies. Second, the chemical environment, namely a monolayer on the surface of a glassy carbon electrode, is unusual compared to the gas phase or the binding pocket of a protein featured in other studies of UCONGA. Third, these molecules have a unique axis, namely the bond to the surface, which makes bounding-box-based analysis different to that performed so far on gas-phase molecules that have no orientational constraints.

Quantifying geometric parameters for these molecules, specifically their height above the surface and their area projected on to the surface, enables calculation of the surface coverage, an important parameter in surface chemistry. Correlating molecular structure with energy is also desirable, because observed molecular properties, such as projected surface areas, are determined as Boltzmann averages over all thermally accessible conformers. Ideally, it would be possible to sort conformers generated by UCONGA into energetically distinct groups before undertaking the time-consuming process of *ab initio* geometry optimization, enabling representative conformers to be selected for subsequent analysis, without losing information on energetic or structural diversity. In this chapter, two different avenues are explored for achieving these outcomes: RMSD-based screening and cluster analysis.

## 6.2 Methods

For each of the five molecules illustrated in Figure 6.1, the UCONGA method described in Chapter 2 was used to generate conformer ensembles. The van der Waals scaling factor used was 0.7 and the smallest step size set to 30°. Enantiomeric conformers were treated as identical and the divide-and-conquer algorithm described in Chapter 3 was not used. Since these conformers are for surface-bound molecules, any conformers where part of the molecule was closer to the surface than the bound atom was rejected.

For molecules with four or more rotatable bonds, an RMSD-based filtering algorithm was used to select a sub-ensemble for subsequent geometry optimization. This algorithm was validated using a system with three rotatable bonds for which the entire conformer ensemble could be optimized. As each conformer was generated, its RMSD to all already-generated conformers was calculated; if any of the RMSD values were less than

1.0 Å then the newly-generated conformer was rejected for being to similar to an already-accepted conformer. Only if it was unique by this metric was it accepted. All generated conformers, divided into accepted and rejected, are available in the electronic appendix

The accepted conformers were optimized at B3LYP/6-31G* [2, 3, 4] as implemented in QCHEM 4.2 [5]. The bond to the surface was capped with a hydrogen atom. No frequency calculations were performed; all stationary points were assumed to be local minima.

Both torsional and bounding-box *k*-means clustering performed, as discussed in Section 2.3.2. In addition, clustering based on the bounding cylinder was also performed. The distance metric here was

$$d_{\text{size}} = \sqrt{(r_1 - r_2)^2 + (h_1 - h_2)^2}$$

where $r_1$ is the radius projected onto the surface and $h_1$ is the height above the surface of the bounding cylinder of conformer 1. The bounding cylinder was used to visualize both the bounding-cylinder and the bounding-box clustering, as visualization in three dimensions is difficult.

## 6.3 Results

### 6.3.1 Simple conformer ensembles: *p*-nitrobenzylamine (NBA) and 1-(ferrocenylmethyl)4-phenyl-1,2,3-triazole (FMPT)

*p*-nitrobenzylamine has two bonds that UCONGA considers rotatable, resulting in an initial ensemble of twelve conformers. On optimization, these converged to one unique conformer (Figure 6.2).

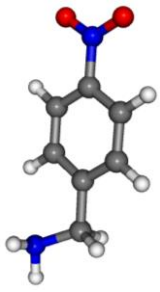


**Figure 6.2** The stable conformer for NBA.

1-(ferrocenylmethyl) 4-phenyl 1, 2, 3-triazole has a third rotatable bond, enough to increase its conformer ensemble to seventy molecules. Upon optimization, these converged to 12 unique conformers. These do not show any meaningful clustering with respect to the torsion angles (Figure 6.3), due to the significant overlap between the clusters.

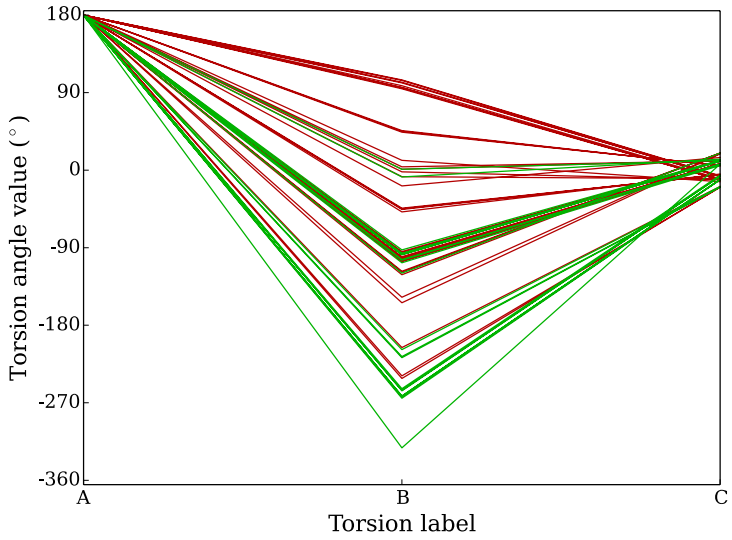



**Figure 6.3** A parallel coordinates plot of the torsion distribution of the optimized conformer ensemble for FMPT. Each line represents a conformer and they are color-coded by cluster identity.

While clustering based on the torsion angles as shown in Figure 6.3 does not identify meaningful differences between conformers within the ensemble, clustering based on the bounding box does. Clustering in three dimensions is hard to visualize, so this was performed in two dimensions using the bounding cylinder instead of the bounding box. A plot of radius against height (Figure 6.4) reveals three bounding-cylinder-based clusters and two bounding-box-based clusters. Both low-radius clusters are low in energy, with all conformers within 2 kJ $mol^{-1}$ of the minimum, but all conformers in the high-radius cluster are between 3.5 and 7 kJ $mol^{-1}$ higher in energy than the minimum. Using the bounding box instead of the bounding cylinder results in the two low-height clusters mostly coalescing. The conformers closest to the bounding-box cluster centers are shown in Figure 6.5.

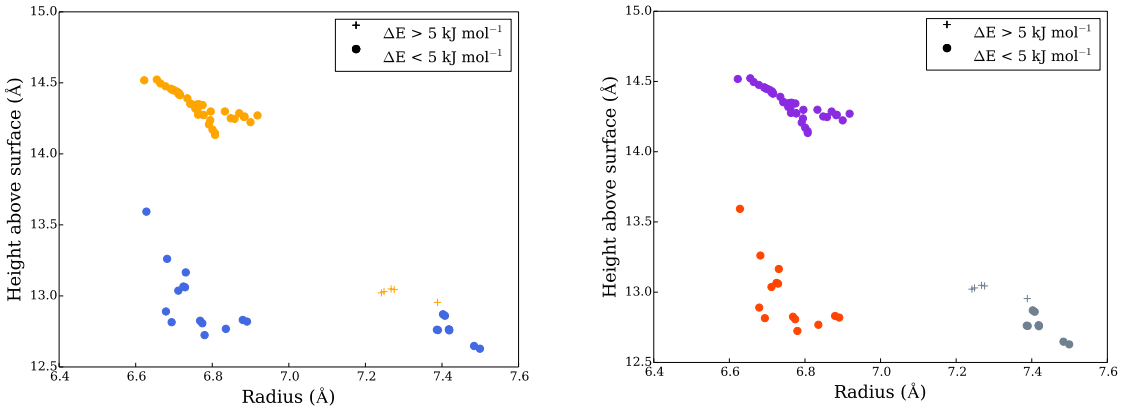



**Figure 6.4** The results of bounding-box based (left) and bounding-cylinder-based (right) cluster analysis for the conformer ensemble of FMPT, visualized in both cases using the bounding cylinder. The color indicates which bounding-box-based cluster the conformer falls into, while the shape of the marker indicates the stability of the conformer compared to the global minimum.

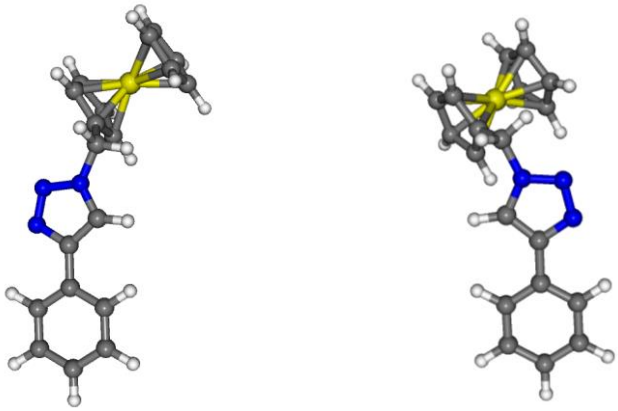


**Figure 6.5** The structures of the closest conformers to the centers of the bounding-box clusters shown in orange (left) and blue (right) in Figure 6.4. The conformer from the orange cluster is more extended than the one from the blue cluster, due to the relative orientation of the ferrocenyl group.

### 6.3.2 Filter validation

The remaining molecules BNB, BFA and FMBA have four or five rotatable bonds. Given the exponential growth in size of conformer ensemble with number of rotatable bonds established in Chapter 2, UCONGA is expected to generate several hundred conformations for them. However, if these large ensembles are analogous to those for NBA and FMPT, many of the generated conformations would converge to the same conformer on geometry optimization. Therefore, it is worth attempting to filter the conformation ensembles generated by UCONGA to remove this duplication before the computationally expensive *ab initio* geometry optimization.

The RMSD-based filtering algorithm described in the methods section (2.3.1) was tested using the conformation ensemble for FMPT, since the conformation ensemble is small enough for all conformers to be optimized but large enough that filtering is meaningful. The test aimed to determine whether or not the filtering changed the average size of the molecule on the surface, where the average is weighted by the population of the conformer. The Boltzmann-weighted average radii and heights are calculated according to:

$$r_{av} = \sum_i \frac{e^{-\Delta E_i / RT}}{\sum_i e^{-\Delta E_i / RT}} r_i$$

where $\Delta E_i$ is the difference between the energy of the conformer and the global minimum of a given conformer ensemble, T = 298K and $r_i$ is the radius of each conformer projected onto the surface. The results reported in Table 6.2 demonstrate that the filtering process identifies a representative ensemble, because the values are similar both pre- and post-filtering. This is confirmed by inspection of the distribution of heights and radii across the unfiltered and filtered conformer ensembles, as illustrated in Figures 6.6 and 6.7.

**Table 6.2** The energy-weighted average properties of the filtered and unfiltered conformer ensemble for FMPT.

| | **Height (Å)** | **Radius (Å)** |
|---|---|---|
| Unfiltered | 13.90 | 6.80 |
| Filtered | 13.92 | 6.79 |

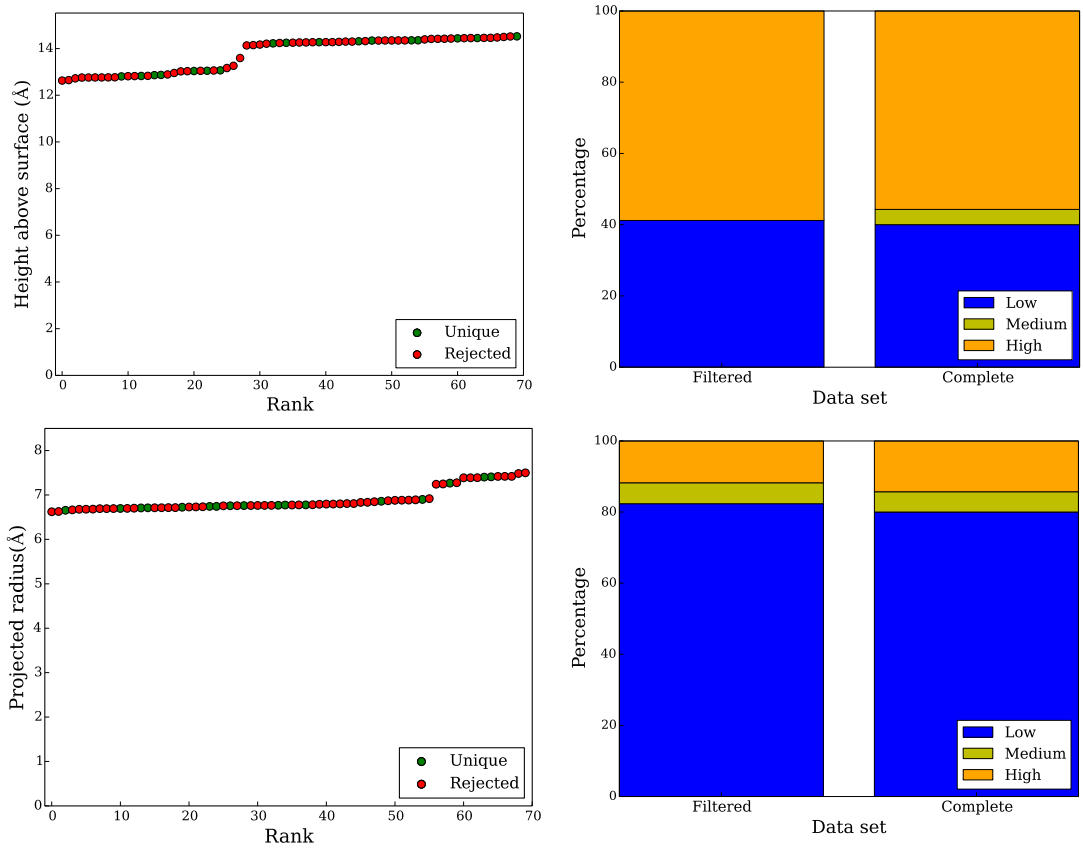



**Figure 6.6** The ensemble distributions (left) and percentages of the ensemble in each cluster (right) for the height above the surface (top) and radius projected onto the surface (bottom) of the total and filtered ensembles for molecule FMPT. Conformers are assigned to different height categories using cutoffs of: low < 13.5 Å, medium < 14 Å, high > 14 Å. Analogous cutoffs for radii are low < 7 Å, medium < 7.35 Å, high > 7.35 Å.

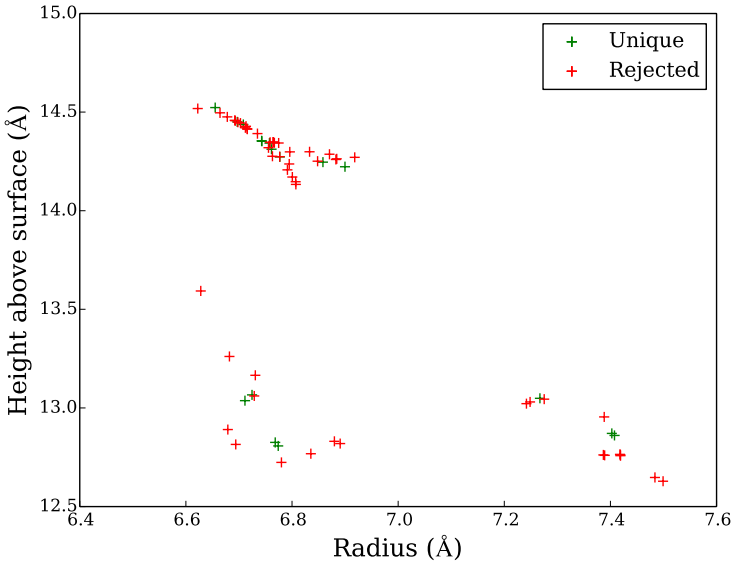



**Figure 6.7** Conformers of FMPT identified as unique during the filtering process are evenly distributed across the molecular size range.

It should be noted that the filtering process excludes the conformer that subsequently optimizes to the global minimum (see Table 6.3). However, as properties are calculated as the Boltzmann average of energetically accessible conformers, this is not a particular problem provided that the remaining conformers form a representative ensemble, as indicated previously.

**Table 6.3** The change in the lowest-energy conformer between the filtered and the unfiltered FMPT ensembles.

| | **Filtered ensemble** | **Unfiltered ensemble** |
|---|---|---|
| Energy of lowest energy conformer (kJ $mol^{-1}$) | 0.51 | 0.00 |
| Radius of lowest energy conformer (Å) | 6.74 | 6.68 |
| Height of lowest energy conformer (Å) | 14.35 | 12.89 |

### 6.3.3 Filtered conformer ensembles: N-benzyl *p*-nitrobenzamide (BNB), N-benzyl ferrocenylacetamide (BFA) and N-(ferrocenylmethyl)benzamide (FMBA)

The complete ensembles generated by UCONGA for BNB, BFA and FMBA contained 1038, 1589 and 285 conformers respectively. Sub-ensembles selected using the filtering algorithm contained 31, 51 and 28 conformers, respectively.

Torsion space clustering does not clearly discriminate between sets of conformers for N-benzyl *p*-nitrobenzamide (Figure 6.8). One cluster (shown in green in the parallel-coordinates plot) contains only 5 conformers and the clusters have significant overlap. It is most likely that the presence of two clusters is simply an artifact of the Calinksi-Harabasz criterion, which does not allow for fewer than two clusters to be generated, dividing by zero to effectively impose an infinite penalty.

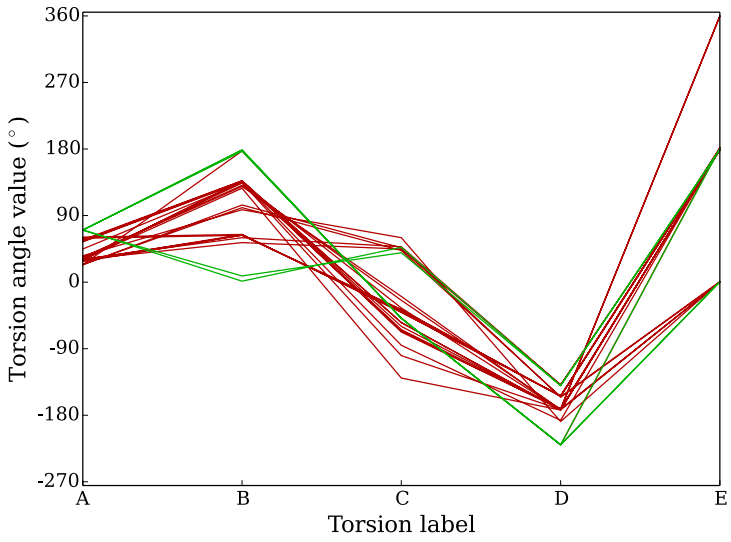



**Figure 6.8** A parallel coordinates plot of the torsion angle distribution for the filtered BNB conformer ensemble. Each line represents a conformer, color-coded by cluster identity.

Bounding-box-based and bounding-cylinder-based clustering both identify two clusters that are somewhat correlated with energetic stability (Figure 6.9). All the high-energy conformers fall into one bounding-box cluster (in blue) and 11 of the 15 low-energy conformers fall into the other. The conformer energies are quite disparate, falling either within 3 kJ mol$^{-1}$ of the global minimum or 20-21 kJ mol$^{-1}$ higher.

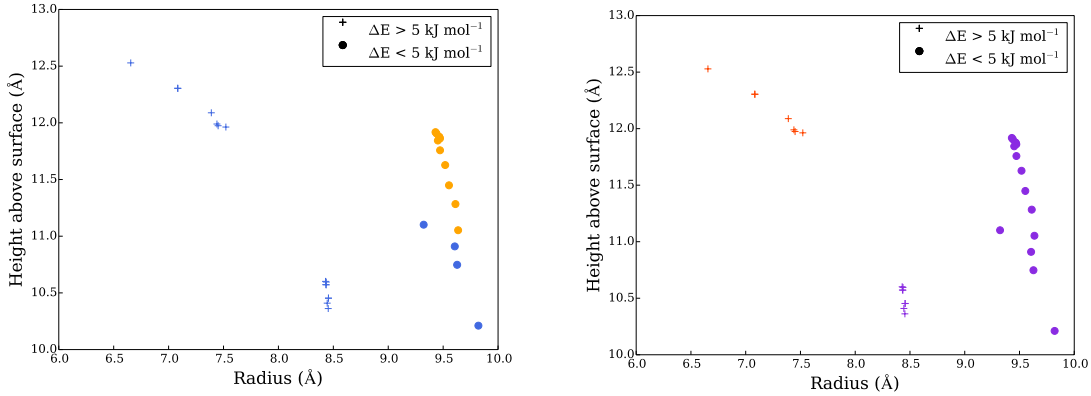



**Figure 6.9** The results of bounding-box-based (left) and bounding-cylinder-based (right) cluster analysis for the conformer ensemble of BNB.

Visual inspection of Figure 6.9 indicates that bounding cylinder based clustering would best correlate with energy, as high radius clusters are uniformly low in energy and low radius clusters uniformly high in energy. If the conformers closest to the bounding-box cluster centers, illustrated in Figure 6.10, are representative, this is because the low-radius

conformers have a steric clash between the amide hydrogen and one of the *ortho* hydrogens.

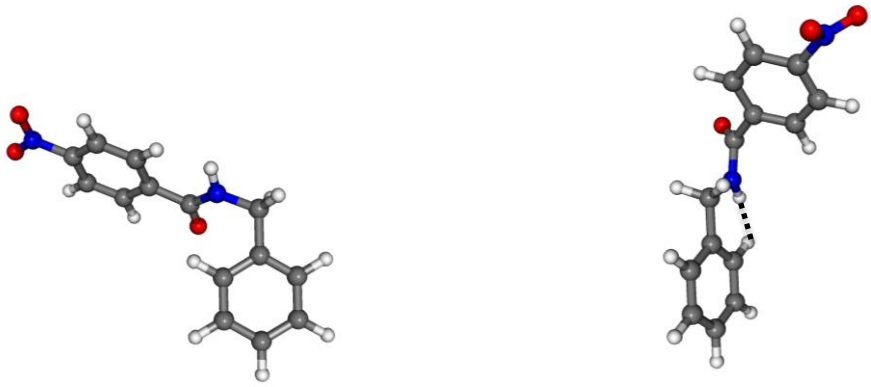


**Figure 6.10** The conformers closest to the center of the low-energy bounding-box cluster shown in orange (left) and high-energy bounding-box cluster shown in blue (right) of BNB, oriented with the surface down and the bond to it capped with hydrogen. The possible steric clash of the high-energy conformer is represented with a dotted line.

In contrast, N-benzyl ferrocenylacetamide (BFA) is weakly clustered based upon molecular dimension (Figure 6.11). Neither bounding-cylinder nor bounding-box clusters are correlated with energy. The torsion-based clusters (Figure 6.12), however, are relatively distinct and correlated to the energy values, with most of the low-energy conformers belonging to the cluster depicted in green in Figure 6.12. Taking the Boltzmann-averaged energies for the clusters as discussed in Section 6.3.2, the red cluster is 5 kJ mol$^{-1}$ higher in energy (inset table). Looking at representative structures (Figure 6.13), this is probably due to greater steric crowding in the red cluster.

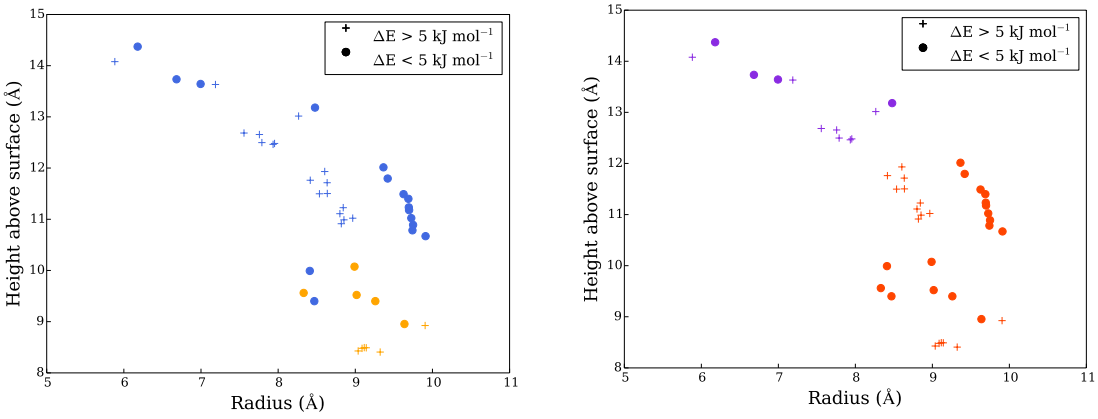



**Figure 6.11:** The results of bounding-box based (left) and bounding-cylinder-based (right) cluster analysis showing poor clustering of the optimized filtered conformer ensemble for BFA.

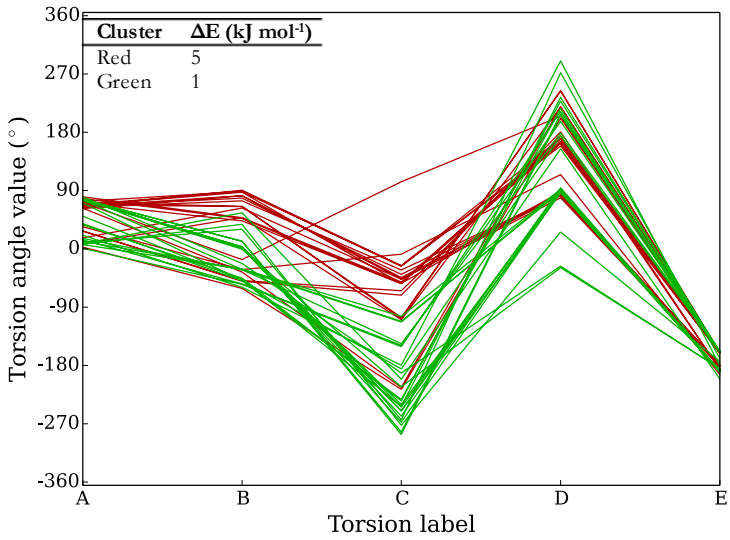



**Figure 6.12:** A parallel coordinates plot (left) showing two reasonably distinct torsion-based clusters. ΔE represents the average energy of conformers in the cluster from the global minimum.

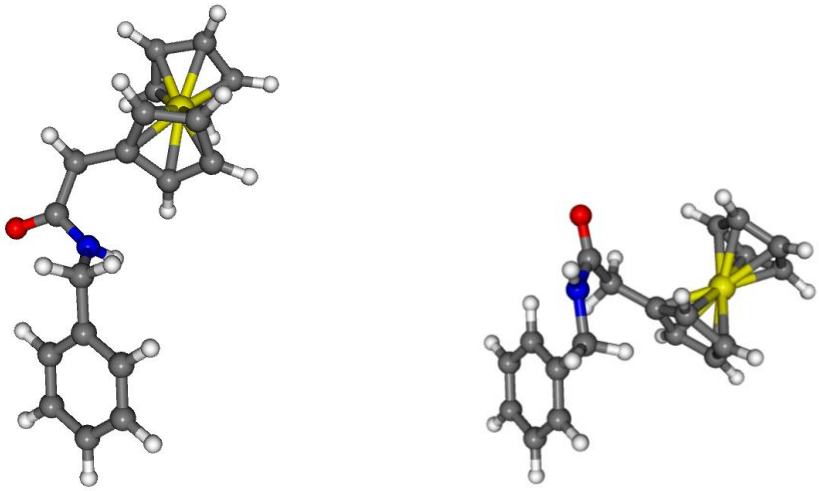


**Figure 6.13:** The structures of the lowest-energy representative molecules from the green (left) and red (right) torsional clusters depicted in Figure 6.12, where representative means that the selected molecules are closest to the cluster centroids. In general, the red cluster contains higher energy conformers than the green cluster.

The torsion-based clustering (Figure 6.14) does not distinguish between conformers within the N-(ferrocenylmethyl) benzamide (FMBA) conformer ensemble, with no visually appreciable difference between clusters, again suggesting that this is an artifact of applying the Calinski-Harabasz criterion discussed previously in this section. Both bounding-box and bounding-cylinder based clustering (Figure 6.15) neatly partition the conformers into extended (high height, low radius, orange) and compact (low height, high radius, blue). The extended conformers are typically lower in energy than the compact ones.

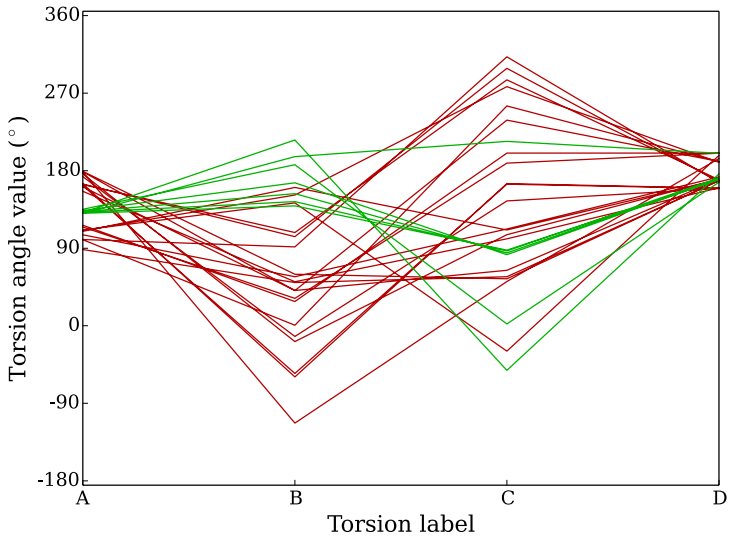



**Figure 6.14:** A parallel coordinates plot showing indistinct torsion-based clusters for the FMBA conformer ensemble.

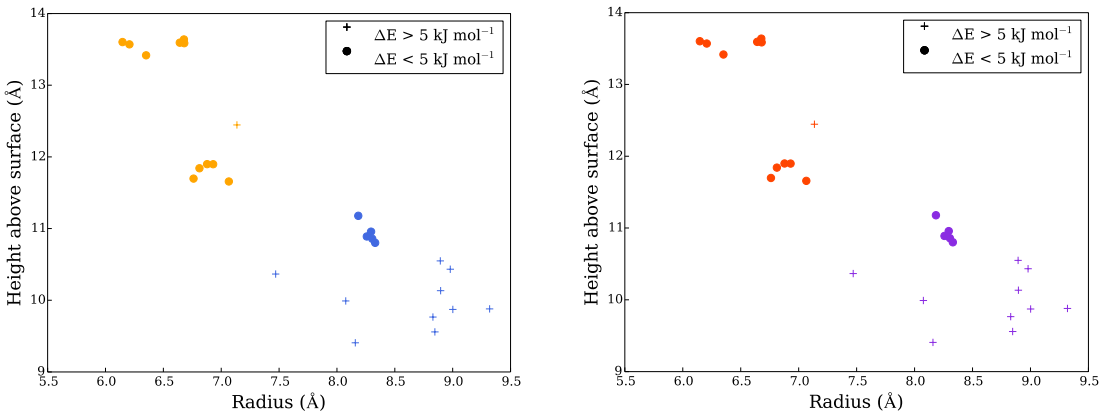



**Figure 6.15:** The results of bounding-box-based (left) and bounding-cylinder-based (right) cluster analysis for FMBA are identical, probably due to the large gap between the two clusters.

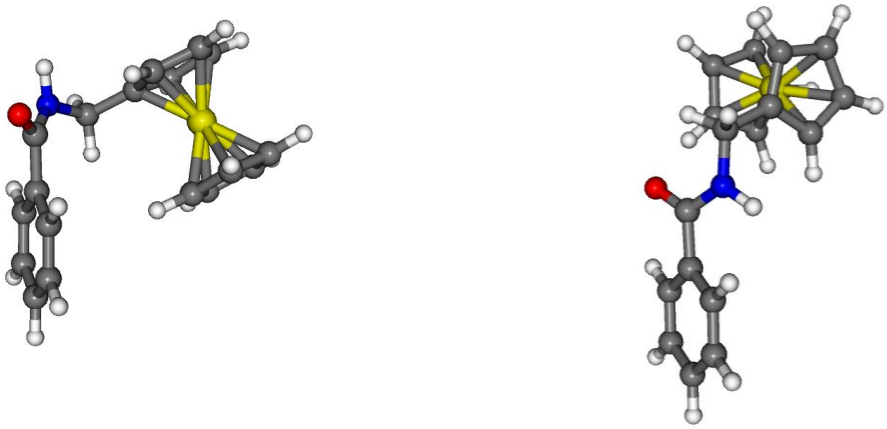


**Figure 6.16:** The structures of the conformers closest to the centroids of the orange (left) and blue (right) bounding-box clusters for FMBA.

## 6.4 Discussion

In general, the two clustering techniques provide complementary information; clustering based on the bounding boxes provides an overall view of molecular shapes within the conformer ensemble, while clustering based upon torsions shows patterns in molecular folding and/or packing. For most molecules studied here, with the exception of BFA, torsional clustering analysis failed to distinguish between conformers, although some torsional structure was observed. Typically, the range of torsion angles accessible by the central bonds within each molecule was restricted, while the bonds at the ends of each molecule were free to rotate, adopting a broader range of torsion angles.

For BFA, two different torsional clusters were observed, one corresponding to mostly extended conformers (green cluster in Figure 6.9, leftmost structure in Figure 6.10) and another to folded conformers, with the terminal bulky groups bridged by a diglycine helix (red cluster in Figure 6.9, rightmost structure in Figure 6.10). These clusters are not correlated with the bounding-box clusters, presumably because the torsional clusters are mostly distinguished by the inner three bonds, while rotation of bulky terminal ferrocenyl group has a dominant effect on the bounding box.

In all cases, the bounding box analysis provided a clear and simple picture of molecular conformations, although did not clearly distinguish between conformers for BFA. Evidently, a wide range of different molecular shapes can be produced by rotating bulky groups on the ends of folded linker units. For the other molecules, with more flexible linkers, it appears that the formation of conformers within the ensemble is driven by the balance between attractive interactions between the terminal groups, producing folded conformers against the torsional strain introduced by folding the linker unit to accommodate this. This leads to a natural separation of compact and extended conformers within the conformer ensemble, as identified by bounding-box clustering.

As a side note, cluster analysis that produces very differently-sized clusters is likely to be unreliable, with clusters produced only as artifact of the Calinski-Harabasz criterion.

Unfortunately, there is not a strong enough correlation between energy and geometry (as measured by either cluster analysis technique) to allow only one or a few representative conformers to be sub-selected from each cluster and further characterized energetically.

To obtain sufficient reliable energy data, as required when calculating Boltzmann-weighted ensemble-averaged properties, it is necessary to characterize the entire filtered conformer ensemble. However, RMSD-based filtering was found to be highly effective in reducing the number of conformers required to form a representative conformer ensemble, without unduly sacrificing accuracy.

## 6.5 Conclusions

The UCONGA algorithm has successfully generated conformer ensembles for moderately flexible molecules tethered to a surface, many of which contain the ferrocenyl moiety. This tethering affects the analysis, as clustering based on the bounding cylinder (which is only possible in an environment where there is one unique dimension, in this case perpendicular to the surface) has been able to detect clusters that the more universal bounding-box clustering has not. In addition, the utility of RMSD-based filtering has been demonstrated. In Chapters 2 and 3, filtering the ensemble using RMSD was primarily an analytical tool to measure the diversity of a conformer ensemble. Here it was used to reduce the size of the ensemble prior to geometry optimization and further analysis. This technique has, in the one case where the generated ensemble was a good size to test the filtering, been found to preserve Boltzmann-weighted average properties and clustering patterns after geometry optimization. Finally, it has been found that no single analysis metric of technique completely describes the conformer ensemble, but that multiple metrics [e.g. atomic coordinates (RMSD); overall conformer size (bounding box and bounding cylinder); similarity between torsion angles and ensemble processing procedures (e.g. filtering, clustering)] are required to extract physical meaning. It can be suggested that overall conformer size is a more easily interpreted metric in all cases and is often the only useful metric for systems with a high degree of torsional flexibility.

## 6.6 References


1) Lee, L.; Gunby, N. R.; Crittenden, D. L.; Downard, A. J. Multifunctional and Stable Monolayers on Carbon: A Simple and Reliable Method for Back Filling Sparse Layers Grafted from Protected Aryldiazonium Ions. *Langmuir* **2016**, *32*, 2626–2637.
2) Stephens, P. J.; Devline, F. J.; Chablowski, C. F.; Frisch, M. J. *Ab Initio* Calculation of Vibrational Absorption and Circular Dichrroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627.
3) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta.* **1973**, *28* (3), 213–222.
4) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665.
5) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, A.; Steele, R. P.; Sundstrom, E. J.; Woodcock III, H. L.; Zimmerman, M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, B.; Brown, S. T.; Casanova, D.; Chang, C.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Jr, R. A. D.; Do, H.; Dutoi, A. D.; Edgar, G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Philipp, H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V; Levchenko, V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.; Marenich, A. V; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; Neill, P. O.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.; Hehre, W. J.; Klamt, A.; Schaefer III, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, D.; Dreuw, A.;

Dunietz, B. D.; Furlani, T. R.; Steven, R.; Hsu, C.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Lyudmila, V.; Subotnik, J. E.; Voorhis, T. Van; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M.; Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *111* (2), 184–215.

6) Caliński, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.* **1974**, *3* (1), 1–27.

# Chapter 7

## Case study: Flexible molecules

## 7.1 Introduction

Generating conformer ensembles for highly flexible molecules is a challenge due to the exponential increase in the number of trial conformers tested, the number of sterically-allowed trial conformers accepted and the number of conformers that stay unique upon geometry optimization. Often many of the local minima on the potential energy surface are close to each other in energy, so the ensemble is large and complicated even after geometry optimization. The use of filtering methods, to automatically reduce the complexity of these ensembles and analysis methods to identify interesting features, is essential for extracting chemical meaning.

When benchmarking UCONGA (Chapter 2), many of the largest conformer ensembles were generated for molecules with flexible rings. This is because changes in ring conformers are less coupled to the rest of the molecule and (unless they lead to a masthead clash or similar) are unlikely to be rejected. Therefore, ring conformers within a molecule tend to have a multiplicative effect on ensemble size, which can only be reduced by pre-screening to remove near-redundant ring conformers, i.e. those that have similar spatial arrangements and do not alter the sterically-allowed conformations of the rest of the molecule. Analyzing ring conformers can be difficult because the overall shapes of the different conformers, as measured by their bounding boxes, are not necessarily sensitive to changes in ring conformations. RMSD is a more sensitive metric in this regard, although may not capture coupling between ring conformations and the conformation of the rest of the molecule. Torsion angle differences, on the other hand, may overstate the difference between conformers. Therefore, studying the generation and analysis of conformer ensembles for molecules containing flexible rings provides a good test of the universality of UCONGA.

A selection of flexible molecules, including molecules with flexible rings and purely aliphatic systems, have been chosen for further study (Figure 7.1). Case studies a-c are drawn from the ASTEX data set, while case study d is of interest as a chemical vapor deposition precursor whose gas phase structure is unknown, and so provides a *de novo* conformer generation test case. Case study a, lysine, is an interesting molecule because, in addition to its protein-bound form in the ASTEX data set, its gas-phase structure is also known. It has complicated gas-phase conformational behavior, with multiple minima close in energy to each other. Case study b, deoxoretinal, contains cyclohexene, a simple

ring system with relatively well-understood conformational properties and a chain of conjugated bonds, a common structural motif. Case study c, dehydroepiandrosterone, contains the steroid ring system, which is more complicated, well-studied and biologically important. Finally case study d, dimeric aluminum isopropoxide, is larger, highly symmetrical, organometallic and contains a branched structure instead of a straight-chain one. While it does feature an $Al_2O_2$ ring, this is planar and so has no other conformers. This combination of a planar core ring with a branched peripheral structure is unique among the molecules studied and adds diversity to this set of molecules.

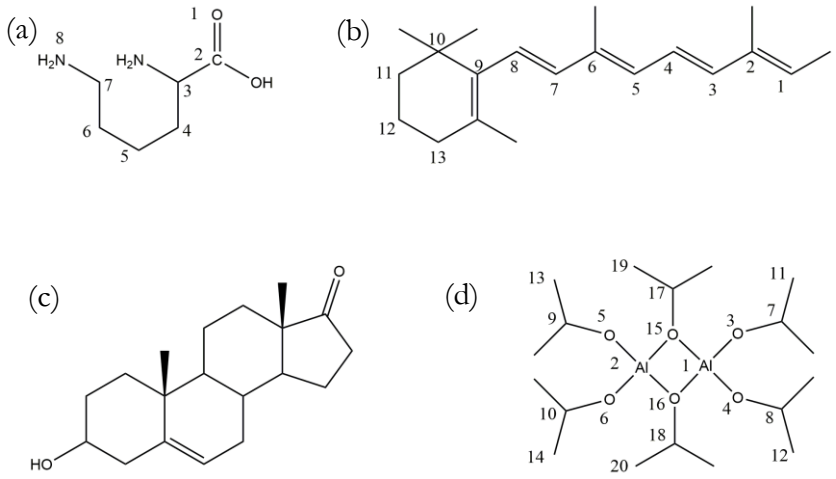



**Figure 7.1** The chemical structures of (a) lysine, (b) deoxoretinal, (c) dehydroepiandrosterone and (d) aluminum isopropoxide dimer, with atoms numbered for systems with multiple rotatable bonds (the bonds themselves are named in Table 7.1).

**Table 7. 1** The bonds labels for molecules with multiple rotatable bonds.

| Lysine | | Deoxoretinal | | Dimeric aluminum isopropoxide | |
|---|---|---|---|---|---|
| 1-2-3-4 | A | 1-2-3-4 | A | 1-15-17-19 | BI1 |
| 2-3-4-5 | B | 3-4-5-6 | B | 2-16-18-20 | BI2 |
| 3-4-5-6 | C | 5-6-7-8 | C | 15-1-3-7 | TO1 |
| 4-5-6-7 | D | 7-8-9-10 | D | 15-1-4-8 | TO2 |
| 5-6-7-8 | E | 10-11-12-13 | E | 16-2-5-9 | TO3 |
| | | | | 16-2-6-10 | TO4 |
| | | | | 1-3-7-11 | TI1 |
| | | | | 1-4-8-12 | TI2 |
| | | | | 2-5-9-13 | TI3 |
| | | | | 2-6-10-14 | TI4 |

## 7.2 Methods

For all the molecules from the ASTEX data set, conformers were created using the UCONGA method with an initial 60° step and a secondary 30° step. All enantiomeric conformers were treated as identical. Due to the greater size of dimeric aluminum

isopropoxide, its conformer ensemble was generated using the divide-and-conquer method and only the 60° step size was used. This was required to generate the conformer ensemble in an acceptable amount of time. Its conformers were optimized at the MP2/6-31G* level of theory [1-3] as implemented in GAMESS [4-5]. The structures of all generated conformers are available in the electronic appendix.

As for Chapter 6, *k*-means cluster analysis was performed, using the Calinski-Harabasz criterion [6] to determine the optimal number of clusters and employing two different distance metrics; one based upon torsion angles and the other on molecular dimensions. The lysine and deoxoretinal ensembles were filtered based on RMSD as for Chapter 6. In addition, the distribution of the RMSDs for each conformer in the ensemble to the reference conformer was used as an additional way to measure the diversity of the ensemble.

## 7.3 Results

### 7.3.1 Lysine

Lysine has five rotatable bonds and relatively little steric hindrance. UCONGA therefore produces a large conformer ensemble, containing 4226 conformers. This large ensemble contains a conformer within 0.18 Å of the experimental reference structure, which is considered a good fit as it is well within the cutoff of 1.0 Å commonly used in the literature [7] and previously in Chapters 2 and 3. On RMSD-based filtering this is reduced to 212 conformers. While the closest conformer to the reference is excluded during filtering, the filtered ensemble nonetheless still contains conformers considered a good fit by RMSD. The closest conformer to the reference within the filtered ensemble has an RMSD of 0.42 Å and is illustrated in Figure 7.2(b). Using RMSD similarities to the reference conformer as a crude measure of diversity (Figure 7.3) filtering reduces diversity slightly, but this is to be expected given the 20-fold reduction in ensemble size. While this suggests that the filtering is representative, it is insufficient to prove it for chemical applications. These require that the Boltzmann-averaged properties, not just the property distributions, are unchanged on filtering. However, the geometry optimization required to confirm this would presumably reduce diversity in the unfiltered ensemble as some conformers converged to each other, and would be quite time-consuming.


**Figure 7.2** The structures of the closest conformers (grey) overlaid on the reference (black) (a) before and (b) after filtering.

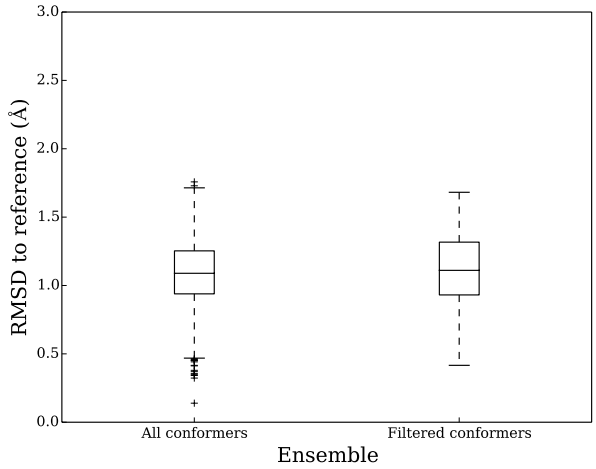



**Figure 7.3** Distribution of the fit to the ASTEX reference for all conformers in both the filtered and unfiltered lysine ensembles.

On torsion-based clustering of the unfiltered ensemble, two clusters of similar sizes are located (Figures 7.4). These differ mainly in the B and D torsion angles, although there is still some overlap, making it hard to extract physical meaning. Clustering based on the bounding box splits the conformer ensemble into two similarly sized clusters that have a small region of overlap regardless of which two axes of the bounding box are used to visualize the ensemble. This suggests that there is no meaningful difference between the bounding-box clusters and that two clusters are produced solely because the Calinski-Harabasz criterion effectively prohibits retention of a single cluster as discussed in Chapter 6.

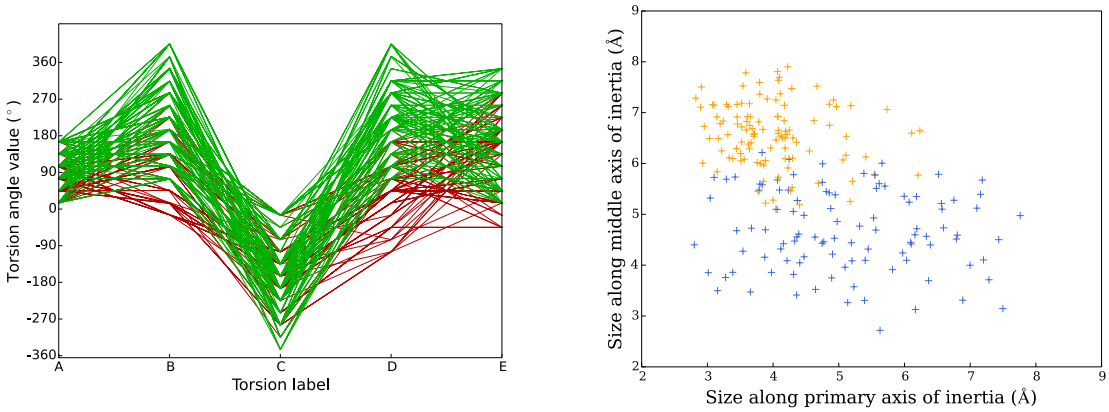



**Figure 7.4** Parallel coordinates plots of the torsion clustering (left, bond labels are in Table 1) and scatter plots of the bounding-box-based clustering (right) of the filtered lysine conformer ensemble. The unfiltered ensemble is similar but due to the size of the ensemble the data points overlap frequently, making the graphs hard to read.

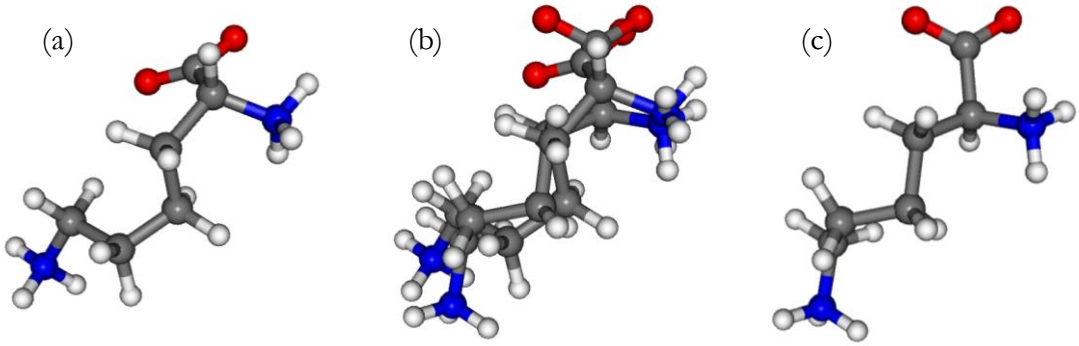



**Figure 7.5** The structures of the closest conformers to the centers of the (a) red and (c) green conformers of lysine, overlaid on each other for comparison (b).

The reference structure for lysine is from a protein crystal structure, where lysine is bound to the protein as a ligand. In this case, lysine is zwitterionic. However, in the gas phase, lysine is uncharged and so adopts different conformations to those observed in the condensed phase. However, as UCONGA is a universal method, it should be able to generate both gas phase and condensed phase conformers.

The gas phase potential energy surface has been explored computationally [8], even rotating around the C-OH and C-$NH_2$ bonds that UCONGA ignores, producing a complex conformer ensemble. The two lowest energy gas-phase conformers are within 1 kJ $mol^{-1}$ of each other and there are 13 other conformers within 7 kJ $mol^{-1}$ of the global minimum. The lowest-energy conformer therefore only constitutes 10% of the ensemble at 298 K. All of the low-energy conformers feature extensive hydrogen-bonding.

The filtered conformer ensemble contains a conformer within 0.86 Å of the lowest-energy gas-phase conformer, which is taken to mean that the lowest-energy gas-phase conformer has been successfully reproduced. The second-lowest energy conformer is not found even in the unfiltered ensemble, where the closest conformer is 1.7 Å away. The reason the second-lowest energy gas-phase conformer is not reproduced by UCONGA is probably due to the two-stage conformer generation process used by UCONGA (Figure 7.6). This involves initially generating trial conformers using a 60° step size, and then altering the rejected trial conformers using a 30° step size. This conformer contains a B torsion angle of 152°, very close to the value of 150° which would be generated in the second stage. However, this would only be generated if the trial conformers with 180° and 120° B torsions generated in the first stage were rejected. Torsion angles closer to a multiple of 30° than a multiple of 60° such as this normally only occur when steric hindrance prevents adoption of the torsion angle close to 60°, but in this case it is preferred due to hydrogen bonding (Figure 7.7), which UCONGA does not account for. The distribution of RMSDs to these reference conformers do not change much on filtering, beyond some of the outliers being removed (Figure 7.8) as for the ASTEX reference conformer.

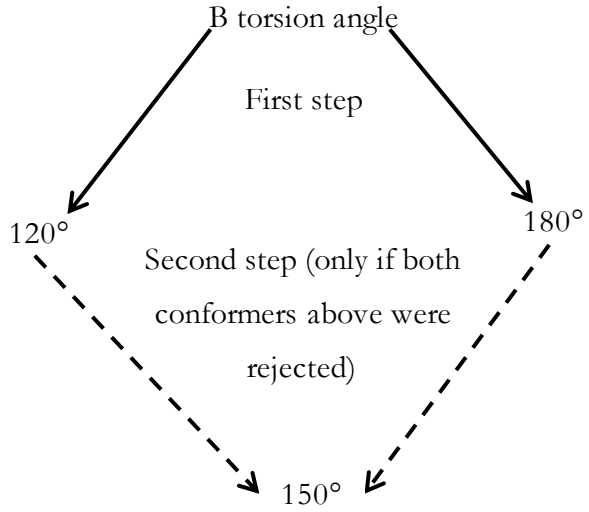



**Figure 7.6** A schematic showing the path by which UCONGA would generate a trial conformer with a 150° torsion angle.

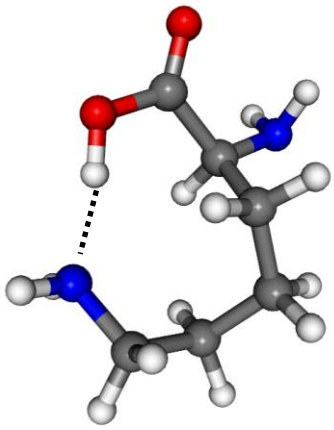


**Figure 7.7** The second-lowest energy gas-phase conformer of lysine, with the hydrogen bond represented as a dashed line.

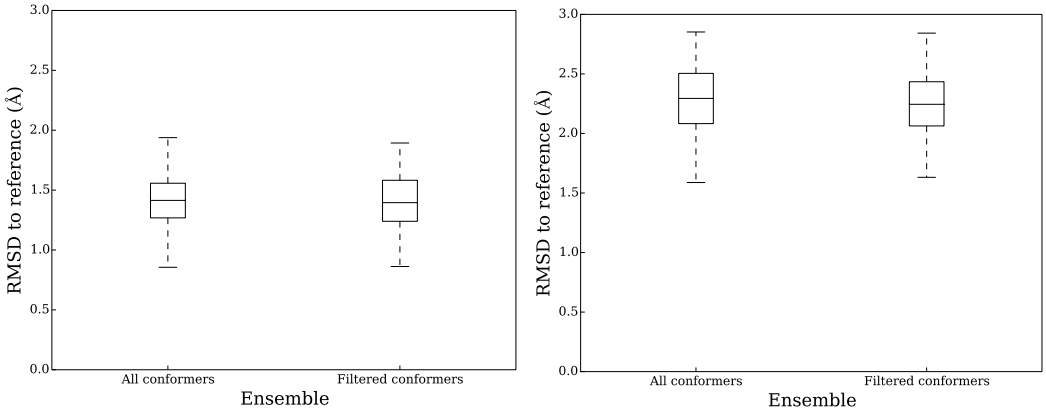



**Figure 7.8** Distribution of fit to the lowest-energy (left) and second-lowest-energy (right) gas-phase reference conformers for all conformers in both the filtered and unfiltered lysine ensembles.

### 7.3.2 Deoxoretinal

Deoxoretinal, like lysine, has a long, relatively unhindered, chain containing a number of bonds classified by UCONGA as rotatable. The resulting unfiltered conformer ensemble is quite large, containing 3240 conformers. One of these conformers has a RMSD of 0.43 Å to the corresponding reference conformer. As with lysine, filtering reduces the ensemble size significantly – to 90 conformers, in this case. The filtered ensemble does not contain the closest conformer to the reference, but it still contains conformers that are a good fit; the closest is within 0.47 Å. Filtering does reduce the diversity slightly, as it did for lysine (Figure 7.9), but loss of some of the outliers is likely inevitable with a 36-fold reduction in ensemble size.


**Figure 7.9** The distributions of the fit to the ASTEX reference for all conformers in both the filtered and unfiltered deoxoretinal ensembles.

Bounding-box clustering identifies two differently sized clusters, distinct in all dimensions (Figure 7.10). The conformers closest to the center of each cluster are shown in Figure 7.11. The larger cluster, shown in blue in Figure 7.10, consists of extended conformations, while the smaller orange cluster is more compact. The ring conformation has no noticeable effect on the bounding box dimensions, and hence each cluster contains a mixture of different ring conformers, with no noticeable correlation between cluster and ring conformer identities.

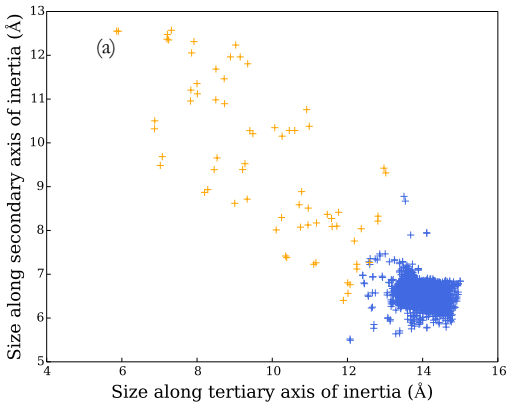



**Figure 7.10** The bounding-box clusters of the deoxoretinal conformer ensemble illustrated using the dimensions associated with the two smallest moments of inertia.

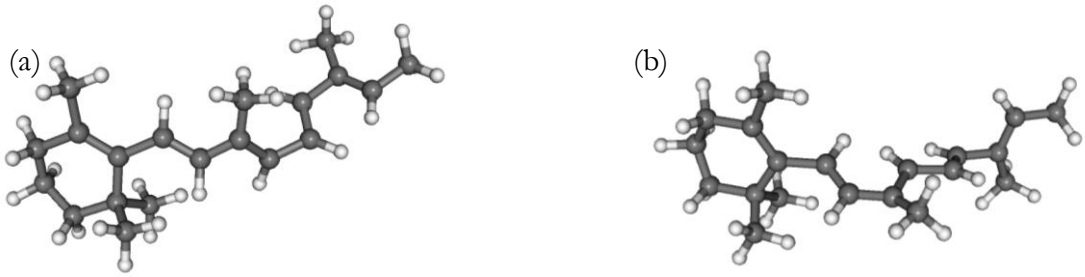



**Figure 7.11** The structures of the conformers closest to the centers of the (a) extended blue and (b) more folded orange clusters of the deoxoretinal conformer ensemble.

Torsion-based clustering also identifies two clusters (Figure 7.12). These are poorly separated, to the point where one of them is buried under the other. Color-coding the bounding-box scatterplot by torsional cluster instead of bounding-box cluster reveals that this clustering causes no variation in bounding box and indeed there are often pairs of molecules with quite similar bounding boxes, one in each cluster. Visualizing the cluster centroids (Figure 7.13) and one of the pairs (Figure 7.14), it seems that the pairs of conformers are diastereomeric, differing in the relative orientation of the ring (E) and chain (A – D) bonds.

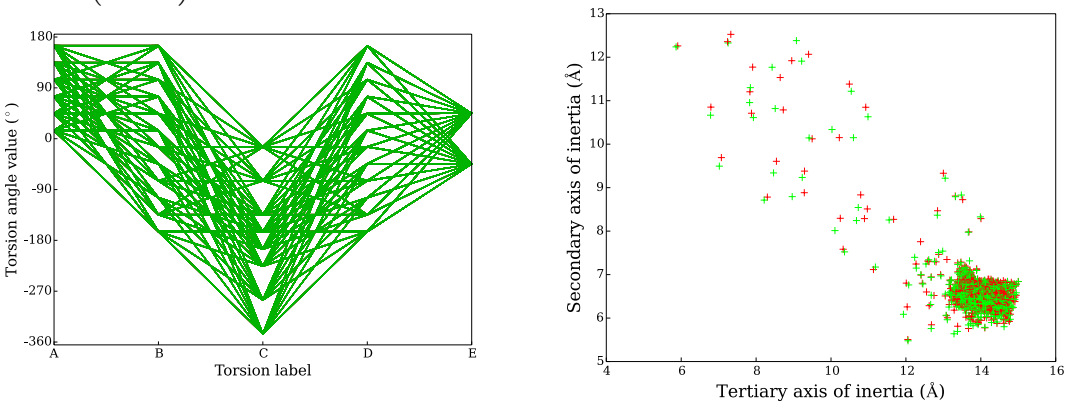



**Figure 7.12** Left: A parallel coordinates plot of the torsion angles of the deoxoretinal conformer ensemble, with conformers colored according to torsional cluster. Right: A scatter plot of the bounding boxes of the deoxoretinal conformer ensemble, with conformers colored according to torsional cluster.

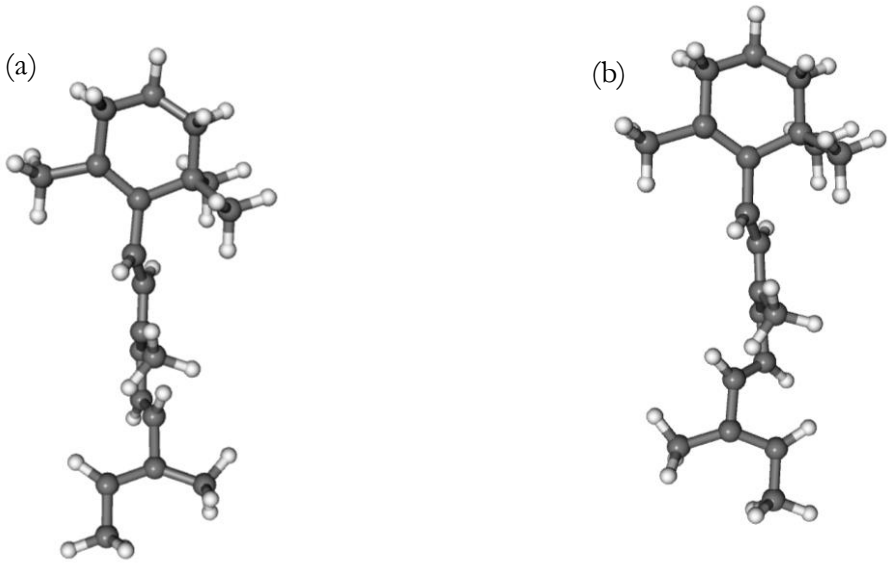



**Figure 7.13** The centers of the (a) green and (b) red torsional clusters have enantiomeric conjugated chains rings but identical rings conformers.

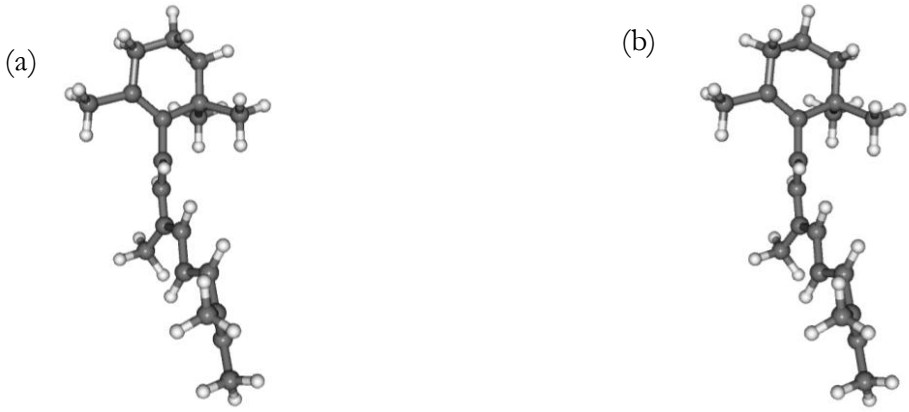



**Figure 7.14** A pair of conformers from the (a) green and (b) red torsional clusters have enantiomeric rings but the same conjugated chains.

From the parallel coordinate plot (Figure 7.12), we see that the bonds in the conjugated chain (A-D) rotate relatively freely. This occurs because UCONGA does not include electronic effects and thus does not penalize breaking conjugation, so many conformers are generated that the system is unlikely to adopt. This inefficiency is a necessary consequence of universality, as electronic effects are less universal than steric ones. On a more positive note, both half-chair forms of the cyclohexene ring (bond E) have been located.

Unlike with lysine, however, there is a significant loss of information on filtering. One of the bounding-box clusters, namely the more compact form shown in orange in Figure

7.10, is discarded entirely (Figure 7.15). The molecular structure of deoxoretinal and the properties of RMSD as a metric suggest a cause. As discussed in Chapter 2, RMSD tends to heavily weight large rigid groups compared to flexible chains. Deoxoretinal, unlike any other molecule filtered using RMSD to date, has a large, relatively rigid group, namely the trimethylcyclohexene ring, on one end of the molecule and nothing on the other. This has caused RMSD to ignore the diversity in the terminus of the conjugated chain, resulting in a lack of correlation between the RMSD between two conformers and the difference in their bounding-box diagonals, leading to the disappearance of the orange cluster with a more tightly-folded conjugated chain.

As a result, the two metrics of similarity are completely uncorrelated. This is shown in Figure 7.16, which compares the structure of the most compact conformers (labeled A on Figure 7.10) with the structure of the most similar accepted conformer. These have sufficient similarity in the ring, and a region partway up the chain, that their differences in the end of the chain are ignored. We expect that similar problems may occur if RMSD-based filtering is used for other molecules where a chain of rotatable bonds has a much larger capping group on one end than it does on the other. The development of a more sophisticated filtering algorithm for these cases may be useful. This could possibly be based on considering the contributions of individual atoms to the RMSD; if many atoms make a small contribution then the conformers are likely to be similar while if (as happens in this case) a few atoms make a large contribution, then the conformers are likely to be different but with a heavily-weighted similar region, as occurs here.

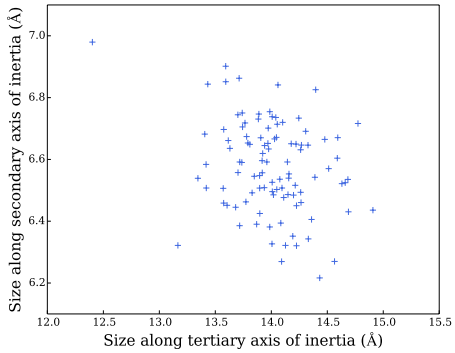



**Figure 7.15** No conformers in the bounding-box cluster depicted in orange in Figure 7.10 are present in the filtered ensemble for deoxoretinal.

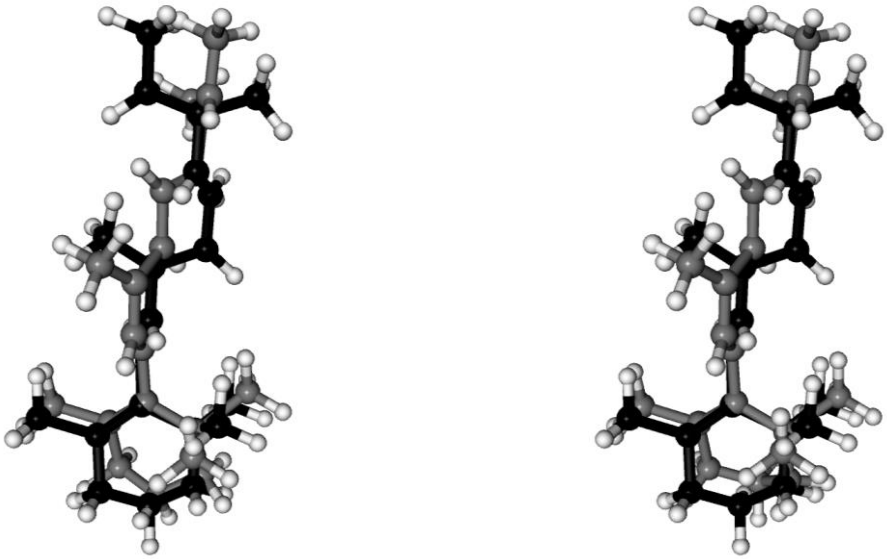


**Figure 7.16** The structures of the two conformers labeled (a) on Figure 7.10 (grey) overlaid on the accepted conformer that caused them to be rejected (black).

### 7.3.3 Dehydroepiandrosterone

For this molecule, only two conformers were generated; the reference conformer and its mirror image (Figure 7.17). Due to the trans ring junctions, the chair conformers cannot ring-flip. The boat form of the A ring, however, is not forbidden. There may be no allowed transition state that can convert the chain form to it and it will almost certainly be higher in energy, but neither of these will be penalized by UCONGA. This boat conformer was not generated, probably because it was rejected for incurring a masthead clash (Figure 7.18).

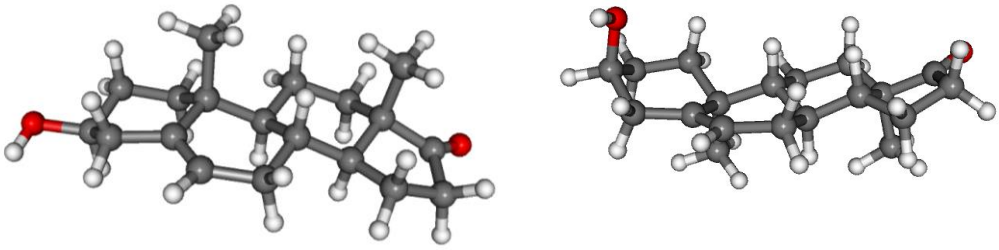


**Figure 7.17** The two conformers of dehydroepiandrosterone which interconvert through a concerted flip of all rings

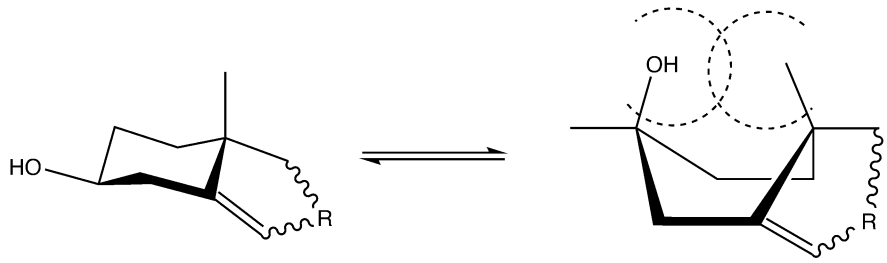



**Figure 7.18** If the A ring of dehydroepiandrosterone were to flip to a boat configuration, the methyl and hydroxyl groups would clash, leading to this trial conformer being rejected.

### 7.3.4 Aluminum isopropoxide dimer

This is the largest system studied in this chapter, with 10 rotatable bonds. However, it has a high degree of nuclear permutational symmetry; the two isopropoxyl groups bonded to each aluminum atom are equivalent, as are the two $Al(O^iPr)_2$ groups. In addition, it is more sterically crowded than lysine or deoxoretinal. Therefore, only 75 conformers are generated, and no filtering was necessary. *Ab initio* geometry optimizations could therefore be performed, so the optimized conformer ensemble could be analyzed, including energetic data. All conformers stay unique on *ab initio* geometry optimization. Many of them are low in energy, with 13 conformers making up 50% of the conformer population at room temperature.

Clustering and graphing the torsion angles (Figure 7.19) reveals two similarly sized clusters. It also shows that the three types of torsion angle behave differently, in a way that makes intuitive chemical sense. The torsions around the non-bridging oxygen-isopropyl groups, labeled TI1-4, rotate semi-freely; they adopt eclipsed and gauche orientations but the torsions are not strongly correlated with the other bonds in this group due to the distance of these terminal isopropyl groups from each other. This shows up on the parallel coordinates plot as crisscrossing lines between the coordinates for these torsions [labelled (a) on Figure 7.19], indicating that many values of one torsion occur in conformers with the same value of another. They are instead correlated with the adjacent aluminum-oxygen bond, labeled TO1-4, as that affects their position relative to the rest of the molecule, indicated by tight bundles of lines from each TI to TO pair, which can be seen on the right-hand half of Figure 7.19. The torsions around these terminal aluminum-oxygen bonds are correlated with each other, as they control the position of the relatively bulky isopropyl groups. This is indicated on the parallel coordinates plot by tight bundles of lines between the coordinates for these torsions, labelled (b) on Figure 7.19. Finally, the bonds from the bridging in-ring oxygens to the

isopropyl groups rotate relatively freely. BI1 rotates over 180° as an artifact of the way UCONGA handles inversion symmetry – it is the first bond in the molecule where one end (namely the isopropyl group) has a mirror plane, so it is restricted to rotate by 180° as other conformers would be enantiomeric (Figure 7.20) – while BI2 rotates over the full 360°.

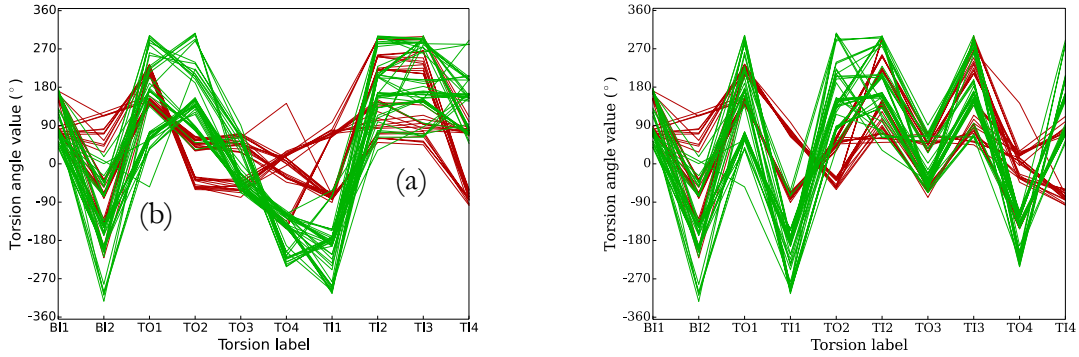



**Figure 7.19** A parallel-coordinates plot of the rotatable bonds of dimeric aluminum isopropoxide, color-coded by torsion-based cluster, with bonds grouped by type (left) showing the correlation of the TO bonds and connectivity (right) showing the correlation of the TO/TI pairs. Bond labels are given in Table 7.1.

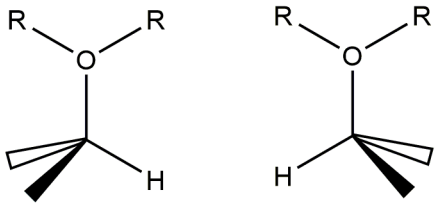



**Figure 7.20** A 180° rotation around the bond to the isopropyl group, combined with interchanging the R group torsions, creates an enantiomeric conformer. Therefore, if rotation around all rotatable bonds in the R groups is free, restricting rotation around the bonds to an isopropyl group is equivalent to avoiding generating enantiomeric conformers.

Clustering based on the bounding box finds two very unevenly sized clusters, despite the tendency of the *k*-means algorithm to produce clusters of similar size. The smaller cluster, of only 7 conformers, has a primary axis length of 9.7-10.7 Å, compared with 8.3-8.9 Å for the rest of the conformers (Figure 7.21). These conformers do not stand out on the parallel coordinates plots of torsion angles, which combined with the structures of the most central conformers of each cluster (Figure 7.22) suggest that it is a collective effect: most conformers have one or two terminal isopropyl groups pointed away from the central Al-O-Al-O ring (that is, the oxygen-isopropyl bond is

approximately aligned with the z axis of this ring), but, for these seven, three or four isopropyl groups are.

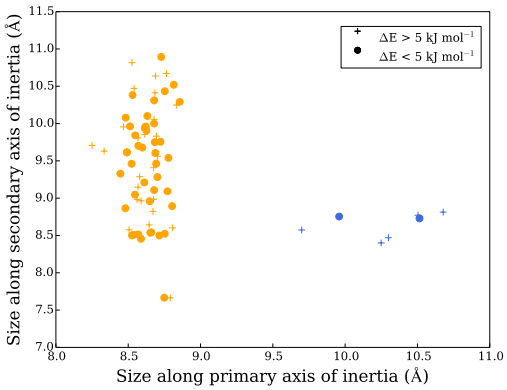



**Figure 7.21** The axes of the two main moments of inertia of the aluminum isopropoxide conformer ensemble, with the color corresponding to bounding-box cluster and symbol to energy.

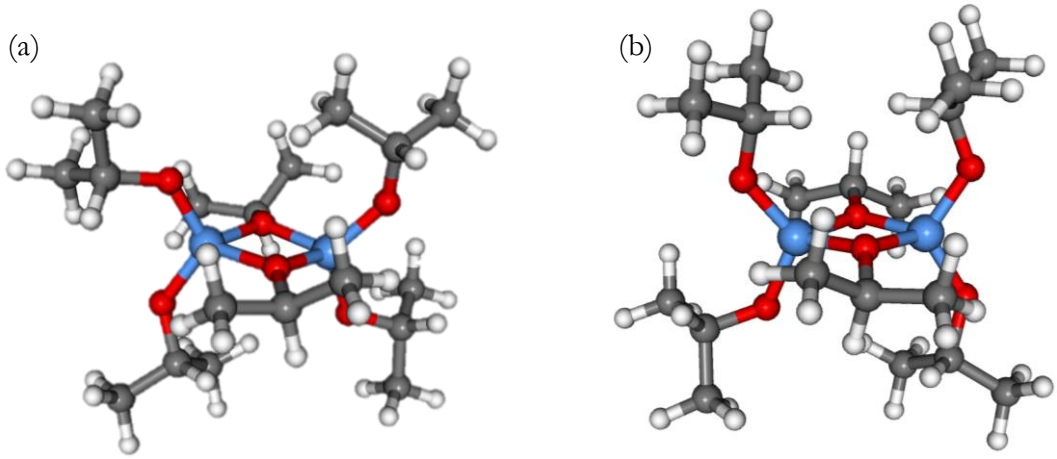



**Figure 7.22** The structures of the conformers closest to the center of (a) the main (in orange on Figure 7.21) and (b) the small (in blue on Figure 7.21) bounding-box-based clusters of the aluminum isopropoxide dimer conformer ensemble. These are aligned so that the primary axis of inertia runs vertically.

## 7.4 Discussion

### 7.4.1 Cyclic systems

Two of the molecules studied have rings that can potentially adopt multiple conformations, one simple ring (the cyclohexene of deoxoretinal) and one more complicated (the steroid ring system of dehydroepiandrosterone). For the first case, UCONGA has successfully reproduced the two known ring conformations, namely the

two half-chair forms. For the second, UCONGA did not produce any other ring conformers, but the only possible one is sterically forbidden.

### 7.4.2 Linear systems

Both lysine and deoxoretinal have very large conformer ensembles due to their lack of steric crowding, allowing free rotation. While RMSD-based screening reduced the ensemble size dramatically, the runtime is long as the conformers have to be generated before they are rejected by RMSD. For such unhindered molecules a conformer ensemble generation method based on steric hindrance such as UCONGA is a sub-optimal choice and knowledge-based methods that can make informed guesses about what sterically allowed conformers will be higher in energy are likely to be more useful. This is especially true for deoxoretinal, where UCONGA breaks conjugation between the double bonds without penalty.

### 7.4.3 Analysis methods

These systems show the important of having complementary analysis techniques. The ability of bounding-box clustering to separate the aluminum isopropoxide dimer and deoxoretinal conformer ensembles into compact and extended clusters would be useful if the condensed phase or supramolecular interactions were being studied. Torsion-based clustering helped classify the torsions based on whether they rotate freely or help separate the clusters for the aluminum isopropoxide dimer and lysine conformer ensembles, which might be useful for further computational investigation. For the aluminum isopropoxide dimer, it was found that the torsions around the nonbridging Al-O bonds are the most important due to their correlations with other torsions, as discussed in Section 7.3.4. For lysine, the B and D torsions separate the clusters, while the A, C and E torsions rotate relatively freely. This information could help perform future studies on these molecules in a more powerful and efficient two-step process. In the first step, the correlated torsions that separate the clusters are focused on, letting the other torsions relax, while in the second step the optimal angles of these torsions are found.

## 7.5 Conclusion

UCONGA has been used to generate conformer ensembles for four flexible molecules, including two containing flexible rings. For all systems, except one gas-phase lysine conformer, a good RMSD of 0.4-0.8 Angstroms to the experimental reference

conformer was achieved, and even that one outlier was reproduced to within a RMSD of 2.0 Angstroms. The utility of having two analysis methods has again been demonstrated, as they have provided different and complementary information. Bounding-box based clustering has located outliers for both the deoxoretinal and aluminum isopropoxide dimer conformer ensembles. Torsion-based clustering, meanwhile, has found the torsions which rotate in a concerted fashion for the lysine and deoxoretinal conformer ensembles, which can direct further studies as torsions which rotate in a concerted fashion can be optimized separately from those the rotate semi-freely. The generated conformer ensembles for these molecules are quite large due to their flexibility and lack of steric crowding. While this makes it more likely that the experimentally relevant conformer will be generated, it increases the time taken to generate and analyze the conformer ensemble. More specialized conformer generation methods may be more efficient in these cases. It was found that RMSD as a metric is ill-suited for use on systems with a large semi-rigid group bonded to a flexible chain with no large group attached to the other end. Finally, it was found that UCONGA is not completely universal; hydrogen-bonding can lead systems to adopt conformers which are not found by UCONGA.

## 7.6 References


1) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222.
2) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77* (7), 3654–3665.
3) Aikens, C. M.; Webb, S. P.; Bell, R. L.; Fletcher, G. D.; Schmidt, M. W.; Gordon, M. S. A Derivation of the Frozen-Orbital Unrestricted Open-Shell and Restricted Closed-Shell Second-Order Perturbation Theory Analytic Gradient Expressions. *Theor. Chim. Acta* **2003**, *110* (4), 233–253.
4) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: GAMESS a Decade Later. In *Theory and Applications of Computational Chemistry*; 2005; pp 1167–1189.
5) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14* (11), 1347–1363.
6) Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.* **1974**, *3* (1), 1–27.
7) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.
8) Leng, Y.; Zhang, M.; Song, C.; Chen, M.; Lin, Z. A Semi-Empirical and *Ab Initio* Combined Approach for the Full Conformational Searches of Gaseous Lysine and Lysine-H2O Complex. *J. Mol. Struct. THEOCHEM* **2008**, *858* (1-3), 52–65.

# Chapter 8

## Conclusions and future work

## 8.1 Conclusions

A family of methods that can generate and analyze conformer ensembles for moderately-sized molecules containing rings and all atom types has been developed. The runtime and generated ensemble size of basic UCONGA scales poorly with the number of rotatable bonds in the molecule so it is generally limited to small molecules of five or fewer rotatable bonds. However, for molecules with favorable structures that reduce the size of the conformer ensemble, especially symmetric bulky branched molecules, up to seven rotatable bonds can be managed. The use of a divide-and-conquer algorithm allows the generation of ensembles for larger molecules, with up to 10 rotatable bonds if they fragment evenly and ring conformers are held fixed.

The strengths and limitations of UCONGA have been illustrated using a series of case studies; bulky molecules, moderately flexible molecules tethered to surfaces and highly flexible molecules. The first limitation is that conformers can be missed where torsion angles and bond angles are coupled. This has been observed for 1,1,2-tri-*tert*-butyldisilane and 1,1,2,2-tetra-*tert*-butyldisilane. The second is that UCONGA, being based purely on steric interactions, does not attempt to preserve conjugation. This has been observed in cases as simple as a nitrophenyl group and as large as deoxoretinal. The third limitation is that conformer ensemble generation is inefficient in both runtime and number of generated conformers for long unbranched chains. Both basic UCONGA and the divide-and-conquer algorithm struggle to prune the conformer ensemble, resulting in a large ensemble containing many highly similar conformers. This has been observed for lysine and, again, deoxoretinal. The last limitation is that RMSD as a measure of similarity weights large rigid or semi-rigid groups more heavily than flexible chains, which can cause problems analyzing molecules consisting of a flexible chain connecting two of these groups, and does cause problems analyzing molecules consisting of a flexible chain attached at one end to such a group. Deoxoretinal is a good example of the latter category.

As for its strengths, UCONGA has been capable of generating conformer ensembles for structurally diverse molecules, including aluminum isopropopxide oligomers, tetrakis(trimethylsilyl)diphosphane, and lysine, with up to 10 rotatable bonds. When these conformer ensembles are too large for *ab initio* geometry optimization to be practical, they can be filtered using RMSD, reducing their size sometimes by 2 orders of

magnitude. Excepting molecules with a flexible chain attached at one end to a bulky semi-rigid or rigid-group, this filtering produces a representative conformer ensemble. Finally, the presence of complementary clustering techniques has allowed the structure in the conformer ensemble to be found for all case studies with sufficiently large ensembles. For very small conformer ensembles – no more than five conformers – visualizing the structures of all the conformers directly is more useful.

RMSD-based filtering, part of the analysis capabilities of UCONGA, can reduce the size of these highly degenerate conformer ensembles by two orders of magnitude while usually maintaining a representative conformer ensemble. The filtered ensemble will not be representative if one end of the chain is attached to a large rigid or semi-rigid group and the other end is not, as RMSD will weight the large group heavily and remove too much diversity from the other end.

## 8.2 Future Work

The main area for improvement in the UCONGA algorithm is in reducing the generated ensemble size. The fact that 10-fold reductions in size frequently occur with RMSD-based screening suggests that many of the conformers generated are too similar to each other. There are three aspects of the UCONGA algorithm that have the potential for improvement to reduce the generated ensemble size: the van der Waals scaling factor, ring conformer generation, and the divide-and-conquer algorithm.

One way to reduce the generated ensemble size is to change the default van der Waals scaling factor. This is currently 0.7, quite a low value. Increasing it would increase the number of trial conformers rejected, which reduces the ensemble size at the risk of rejecting some desirable trial conformers. For most molecules, this is a good tradeoff as there are many accepted trial conformers for each basin on the potential energy surface, all of which would converge to one conformer upon geometry optimization. However, for extremely sterically crowded molecules, there is a higher risk of some basins in the potential energy surface missed entirely. One possible improvement in this area is to select the van der Waals scaling factor based on the steric hindrance of the molecule, possibly calculated using the average number of non-hydrogen neighbors each atom has. For simple linear molecules, with a low number of heavy neighbors, a higher scaling factor would be chosen than for a crowded highly branched system with many heavy

neighbors. Another possible improvement would be to use a greater scaling factor for those trial conformers that are similar to an already-accepted trial conformer, although efficiently determining if a similar conformer had already been generated may be a challenge. In theory, this would mean that a conformer similar to an already-accepted trial conformer would only be accepted if it was less sterically crowded, and hopefully lower in energy. However, the most straightforward implementation of this would involve looping through the list of already-generated conformers for each trial conformer, which would significantly increase the runtime.

There are other aspects of the algorithm that could be improved. Currently, ring conformers make a large contribution to ensemble size as discussed in Chapter 3. This could be reduced for well-studied systems (especially cyclohexane) by trialing only common conformers (such as the two chair conformers for cyclohexane) first and only trying less common conformers (such as twist-boat conformers) if no common conformers are accepted. The current procedure would be used for rings where the conformational properties are not well-studied. However, even among ring systems with heavily studied conformational properties, the utility of this is limited; cyclopentane rings have many low-energy conformers and would not work with this strategy as well as cyclohexane does.

The most obvious place to improve the divide-and-conquer module is its division method. Merging adjacent fragments, such as those separated by a double bond or in a 1,2 relationship in a ring, if the combined fragment is not too large would probably increase coupling and hence the number of fragment trial conformers rejected. This is a tradeoff, though, as it would increase the time spent during fragment conformer generation. Additionally, a larger van der Waals scaling factor could be used for fragment conformer generation than for molecular conformer generation, as the fragments should be less sterically crowded.

In addition to improving the UCONGA algorithm, there are some questions raised by the *ab initio* benchmarking that would be interesting to explore further. Recalculating the M06 energies with a larger integration grid and basis set would be a good starting point towards understanding why the calculated relative energies were so different to the MP2/6-31G* benchmark. Comparing M06 relative energies between program packages would also help verify each implementation of this complex, highly parameterized

functional. In addition, repeating both the B3LYP and M06 calculations using a dispersion corrected functional is likely to improve the accuracy of the calculated relative energies, especially for B3LYP.

# Appendix 1

## Publications, conferences, achievements, service and funding

## A1.1 – List of publications

Gunby, N. R.; Krumdieck, S.; Murthy, H.; Masters, S. L.; Miya, S. S. Study of Precursor Chemistry and Solvent Systems in PP-MOCVD Processing with Alumina Case Study. *Phys. Status Solidi* **2015**, *212* (7), 1519–1526.

Lee, L.; Gunby, N. R.; Crittenden, D. L.; Downard, A. J. Multifunctional and Stable Monolayers on Carbon: A Simple and Reliable Method for Back Filling Sparse Layers Grafted from Protected Aryldiazonium Ions. *Langmuir* **2016**, *32*, 2626–2637.

## A1.2 – Conference contributions

Gunby, N. R.; Masters, S. L. Chemical Studies of a Novel CVD Technique. *New Zealand Institute of Chemistry (NZIC) Conference,* Wellington, New Zealand, **2013**, Poster presentation.

Gunby, N. R. Chemical Studies of a New CVD Technique. *Univerisity of Cantberury Postgraduate Student Showcase,* Christchurch, New Zealand, **2013**, Oral presentation.

Gunby, N. R.; Masters, S. L. Diffraction and Simulation Studies of a Novel CVD Process. *Austin Symposium on Molecular Dynamics @ Dallas,* Dallas, USA, **2014**, Poster presentation.

Gunby, N. R.; Masters, S. L. Computational and Experimental Investigation of Precursors to Pulsed-Pressure Metal-Organic Chemical Vapour Deposition. *Quantum and Computational Chemistry Student Conference (QUACCS),* Cass, New Zealand, **2014**, Oral presentation.

Gunby, N. R.; Masters, S. L. Molecular dynamics studies of flash evaporation of dilute solutions for CVD processes. *WATOC Conference,* Santiago, Chile, **2014**, Poster presentation.

Gunby, N. R. UCONGA: Forcefield-free Universal CONformer Generation and Analysis. *Biomolecular Interaction Center Symposium,* Christchurch, New Zealand, **2014**, Oral presentation.

Gunby, N. R.; Masters, S. L.; Crittenden, D. C. Predicting Molecular Shape. *University of Canterbury Postgraduate Student Showcase,* Christchurch, New Zealand, **2014**, Oral presentation.

Gunby, N. R.; Masters, S. L.; Crittenden, D. L. Conformer generation for gas-phase structural chemistry. *Royal Australian Chemical Institute Physical Chemistry Student Conference,* Adelaide, Australia, **2014**, Oral presentation.

Gunby, N. R.; Masters, S. L.; Crittenden, D. C. UCONGA: Universal Conformer Generation and Analaysis. *University of Canterbury Chemistry Postgraduate Student Showcase,* Christchurch, New Zealand, **2015**, Oral presentation.

Gunby, N. R.; Masters, S. L.; Crittenden, D. L. UCONGA: A new, general, fast conformer generation method. *European Symposium on Gas-phase Electron Diffraction,* Chiemsee, Germany, **2015**, Oral presentation.

Gunby, N. R.; Masters, S. L.; Crittenden, D. C. Predicting Molecular Shape. *University of Canterbury Postgraduate Student Showcase,* Christchurch, New Zealand, **2015**, Oral presentation. (RSC prize for best presentation)

Gunby, N. R.; Masters, S. L.; Crittenden, D. L. UCONGA: A new, general, fast conformer generation method. *Royal Australian Chemical Institute Physical Chemistry Division Meeting,* Christchurch, New Zealand, **2016**, Oral presentation.

## A1.3 – Achievements, professional membership and service

### A1.3.1 Achievements

2015 Co-awarded the Royal Society of Chemistry prize for best presentation at the Department of Chemistry Postgraduate Student Research Showcase

### A1.3.2 Professional Memberships

2013-Present New Zealand Institute of Chemistry (NZIC) student member

### A1.3.3 Voluntary Service

2015 University of Canterbury Postgraduate Students Assocation (PGSA) Treasurer
2014 Nitrate testing for University of Canterbury Outreach program
2014 Burnside Primary School visit for University of Canterbury Outreach program
2014 PGSA Communications Officer

### A1.3.4 Paid Service

2012-2015 Lab supervising/demonstrating for Chemistry undergraduate labs (CHEM111, CHEM112)

## A1.4 – Funding received

| | |
|---|---|
| 2014 | Evans Fund |
| 2014 | Lord Rutherford Memorial Research Fellowship |
| 2014 | Todd Foundation Award for Excellence |
| 2014 | Professor Jim Coxon Graduate Prize in Chemistry |
| 2013 | Canterbury Scholarship |